ANALYSIS OF ITEM DIFFICULTY AND CHANGE IN MATHEMATICAL AHCIEVEMENT FROM 6TH TO 8TH GRADE'S LONGITUDINAL DATA

A Dissertation
Presented to
The Academic Faculty

By

Mee-Ae Kim-O (Mia)

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology

August, 2011

ANALYSIS OF ITEM DIFFICULTY AND CHANGE IN MATHEMATICAL AHCIEVEMENT FROM 6TH TO 8TH GRADE'S LONGITUDINAL DATA

Approved by:

Dr. Susan E. Embretson, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Francis (Frank) T. Durso
School of Psychology
Georgia Institute of Technology

Dr. Lawrence R. James
School of Psychology
Georgia Institute of Technology

Dr. Howard A. Rollins, Jr
School of Psychology
Georgia Institute of Technology

Dr. Joseph L. A. Hughes
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: June 24, 2011

DEDICATION

To Lovely God, Jesus Christ, and the Holy Spirit, who have made all this possible.

"Seek first the Kingdom of God and His righteousness, and all these things shall be added to you." Matthew 6:33

"Trust in the Lord with all your heart; Do not defend on your own understanding. Seek His will in all you do, And He will show you which path to take." Proverbs 3:5-6.

ACKNOWLEDGEMENTS

I wish to thank Dr. Susan Embretson for her great teaching and wonderful guidance as my advisor. I would also like to thank rest of my committee members: Dr. Larry James, Dr. Frank Durso, Dr. Howard Rollins, and Dr. Joseph Hughes. Their assistance has been invaluable and amazing. Also, I can't forget Dr. Greg Corso and Jan Westbrook who helped me abundantly during my Georgia Tech school times. My Wieuca Road Baptist Church members deserve a special acknowledgement for praying as well. I also really appreciate my lovely parents-in law and my wonderful mom. Even though my dad (Michael O) passed away when I started to study in Georgia Tech, I accomplished his dream! He gave me a high motivation to study in the United States when I was young.

I especially owe my lovely husband John H.S. Kim a great debt for his dedication to my studying and living as my co-worker, the best spiritual friend, and husband. A special thanks to my two lovely beautiful daughters, Uni Sarah Kim and Minji Victoria Kim. Finally, I love you all and thank you so much!!! However, my deepest gratitude goes to my amazing God, Jesus Christ, and the Holy Spirit!!! Because of God's helping, caring, and loving, my husband and I, we are DONE!!! Since 1991, I have prayed for getting a Ph.D in the United States. God gave the chance for studying in MTSU and Georgia Tech. Finally, I realized that God heard my prayer! HALLELUJAH!!! During my studying, I have met a lot of difficulties, problems, and obstacles. However, God always gave me wisdom, knowledge, and the understanding ability to overcome them. The kingdom of God and His righteousness will be another goal for me!

THANK YOU!!!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Mathematics is an increasingly important aspect of education because of its central role in technology (Kuenzi, 2008). Mathematical achievement tests are universally applied throughout schooling in the US to assess yearly progress. The middle school years (e.g., Grade 6-Grade8) are especially crucial to success in mathematics because students must acquire the skills needed in Algebra and higher levels of mathematics (National Mathematics Advisory Panel, 2008). The middle school years are also important developmentally because complex reasoning also emerges (e.g., Piaget, Vygotsky) and possibly at different rates for different students. According to many perspectives, the best design for studying changes in achievement and thinking in the middle school years is a longitudinal study of representative samples of children.

For the current study, item responses to mathematical achievement tests administered during the middle school years were available for a randomly selected sample of 2,667 students in a Midwestern state. Until recently, however, inferences from such data were limited by the psychometric methods that were available to scale the data and provide meaningful comparisons. For the current study, some very recent advances in item response theory (IRT) were applied to provide inferences about growth. These methods consisted of confirmatory multidimensional and longitudinal models that previously were impractical to apply to large numbers of items and examinees.

Growth in mathematical achievement was studied in the four major areas covered by the test (Number, Algebra, Geometry and Data) and in some specific areas that were especially consistent in definition across the grades. Differences in growth were also

studied in two areas of individual differences, gender and socio-economic background, that have often been found important in careers that involve mathematics (Kuenzi, 2008). The analyses were conducted in the context of a series of hypotheses about growth and the substantive nature of differences across middle school.

In Study 1, the substantive nature of change over middle school was examined by comparing the strength of the specific content areas across grades. Confirmatory multidimensional IRT models were applied to test hypotheses about concept structures in mathematics. In Study 2, growth was examined by fitting longitudinal IRT models to items from the various content areas. It was found that the relative strength of the content areas shifted somewhat across grades in defining mathematical achievement. The largest growth occurred from Grade 6 to Grade 7. The specific pattern of growth varied substantially by the socio-economic status of the student but few differences emerged by gender. The implications of the results for education and for developmental theories of cognitive complexity are discussed.

CHAPTER 1

INTRODUCTION

1.1 Mathematical Achievement and Middle School Students

Mathematics is an increasingly important aspect of education because of its central role in technology (Kuenzi, 2008). A mathematical achievement test was compared with eighth-grade students across a sample of twenty countries including England, Germany, Canada, Japan, Korea, and the United States. Tatsuoka, Corter, and Tatsuoka (2004) analyzed achievement test data from the Trend in International Mathematics and Science Study-Revised (TIMSS-R) 1999 in order to compare students' knowledge of mathematics and science in those countries. By observing the skill structure of the items on the TIMSS-R, Tatsuoka et al. found clear differences among the countries in the pattern of sub-skill achievement. U.S. students were strong in some content and quantitative reading skills, but weak in many other areas such as algebra, statistics, and, especially, geometry. Therefore, the mathematical achievement of children has been one of the important issues in the U.S. schools in order to maintain global leadership in the future.

Learning mathematics at the elementary and middle school levels forms the basis for achievement in high school and college mathematics, and for the broad range of mathematical skills used in the workplace. Mathematical achievement tests are universally applied throughout schooling in the U.S. to assess yearly progress. The middle school years (e.g., Grade 6-Grade 8) are especially crucial to success in mathematics because students must acquire the skills needed in Algebra and higher levels

of mathematics (National Mathematics Advisory Panel, 2008). The middle school years are also important developmentally because complex reasoning also emerges (e.g., Piaget, Vygotsky) and possibly at different rates for different students. As proposed by Jean Piaget proposed, cognitive development proceeds through four stages: Sensorimotor stage (birth until age 2), Preoperational stage (age 2 until age 6 or 7), Concrete operational stage (age 6 or 7 until age 11 or 12), and formal operational stage (age 11 or 12 through adulthood) (Piaget & Inhelder, 1969). Piaget insisted that children cannot learn from an experience until they have begun the transition into a stage that allows them to deal with and conceptualize that experience appropriately. Vygotsky's concept was that children can be improved when they have the assistance of the more advanced and competent people than themselves.

Gender and socio-economic background have often been found important in careers that involve mathematics (Kuenzi, 2008). First, much research has been done on the gender issue and showed that males outperformed females in mathematical achievement in the past (Hopkins, 2004). However, a majority of the current studies conclude that the gender gap in mathematics has been decreasing in recent decades and is now quite small (Fennema, 1996; Gray, 1996; Hanna, 2003; Leahy & Guo, 2001; Wellesley College, 1992). National Mathematics Advisory Panel (NMAP; 2008) reported that the average mathematical achievement of boys and girls showed very similar levels (in favor of boys) in large nationally representative samples. Second, it is also a well known issue that low SES is linked to poor mathematical performance (NMAP, 2008). Based on the NAEP (National Assessment of Educational Progress) data in 2001 and 2002, the achievement gaps between the high SES and low SES students existed

regardless of race (Hopkins, 2004). SES may have many aspects of definition. NMAP (2008) included parental education, poverty level, parental income, or a composite index to define SES. Therefore, SES differences may be used as a supporting evidence of the Vygotsky's theory that children can improve their school achievement when mentoring is available.

1.2 Objectives of the Present Study

The purpose of the current study was to measure and interpret the mathematical achievement growth during the middle school years using a longitudinal data of the mathematical achievement tests administered from 6^{th} grade to 8^{th} grade. Until recently, however, inferences from such data were limited by the psychometric methods that were available to scale the data and provide meaningful comparisons. For the current study, some very recent advances in item response theory (IRT) were applied to provide inferences about growth. These methods consisted of confirmatory multidimensional and longitudinal models that previously were impractical to apply to large numbers of items and examinees.

Study 1 was conducted as a preliminary study to examine the general properties of the mathematical achievement test before measuring students' achievement change in Study 2. Item difficulties of the mathematical standards and their benchmarks were carefully examined. Also, confirmatory multidimensional IRT models were applied to test hypotheses about concept structures in mathematics.

In Study 2, growth in the mathematical achievement over the middle school years was examined using longitudinal IRT models in the various content areas. The

achievement change was measured and analyzed to see if and where a significant incremental difference exists in mathematical achievement in the middle school years (e.g., between 6^{th} grade and the next two grade levels, 7^{th} and 8^{th}). It was also examined if gender and socioeconomic status (SES) affected the mathematical achievement within and between grades.

CHAPTER 2

THEORETICAL BACKGROUNDS FOR ITEM DIFFICULTY AND COGNITIVE DEVELOPMENT & MATHEMATICAL ACHIEVEMENT

2.1 Understanding Mathematical Item Difficulty

Understanding item difficulty is one of the primary issues in the mathematical achievement tests; identifying the difficulty of the mathematical achievement tests is fundamental and important for measuring the mathematical achievement change. Recognizing the sources of item difficulty and cognitive complexity is useful for better understanding the cognitive requirements in the test as well as for predicting item difficulties using psychometric models. Results from psychometric modeling of item difficulty can provide evidence of construct representation (Embretson, 1983) for the test, can guide test development by defining cognitive variables for item design, and can provide a basis for item banking and automatic item generation (Embretson, 1999).

Factors that motivate the difficulty of mathematics test items have been studied by several researchers. Cognitive complexity and depth of knowledge are considered important aspects of understanding mathematical item difficulty (Embretson & Daniel, 2008). Generally, item difficulty is affected by the cognitive complexity- the cognitive demand to solve an item. Webb (1999) proposed a framework called "Depth of Knowledge" to classify items by their levels of cognitive complexity. This framework is widely used for selecting school achievement test items for many state year-end tests and National tests (Embretson & Daniel, 2008). Four levels of depth of knowledge are as followings:

Level 1: Recall

Level 1 is the recall of information such as a fact, definition, term, or a simple procedure. A sample item at this level is “Determine the perimeter or area of rectangles given a drawing or labels.”

Level 2: Skill/Concept

Level 2 involves the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is more complex than in level 1. For instance, “Compare rectangle and square.” To compare two objects requires identifying characteristics of the objects and then grouping or ordering the objects by those characteristics.

Level 3: Strategic Thinking

Level 3 includes reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity does not result only from the fact that there could be multiple answers, a possibility for both levels 1 and 2, but also from the multi-step task which requires more demanding reasoning than level 1 and 2. It generally takes less than 10 minutes to do. An example for level 3 is solving a multiple-step problem and providing support, with a mathematical explanation that justifies the answer.

Level 4: Extended Thinking

The tasks of level 4 require high cognitive demands and are very complex (e.g., investigation and time to think and process multiple conditions of the problem or task.) Students are required to make several connections to relate ideas within the content area or among content areas and they have to select or devise one approach, among many

alternative approaches, for how the situation could be solved. It takes more than 10 minutes to do non-routine manipulations.

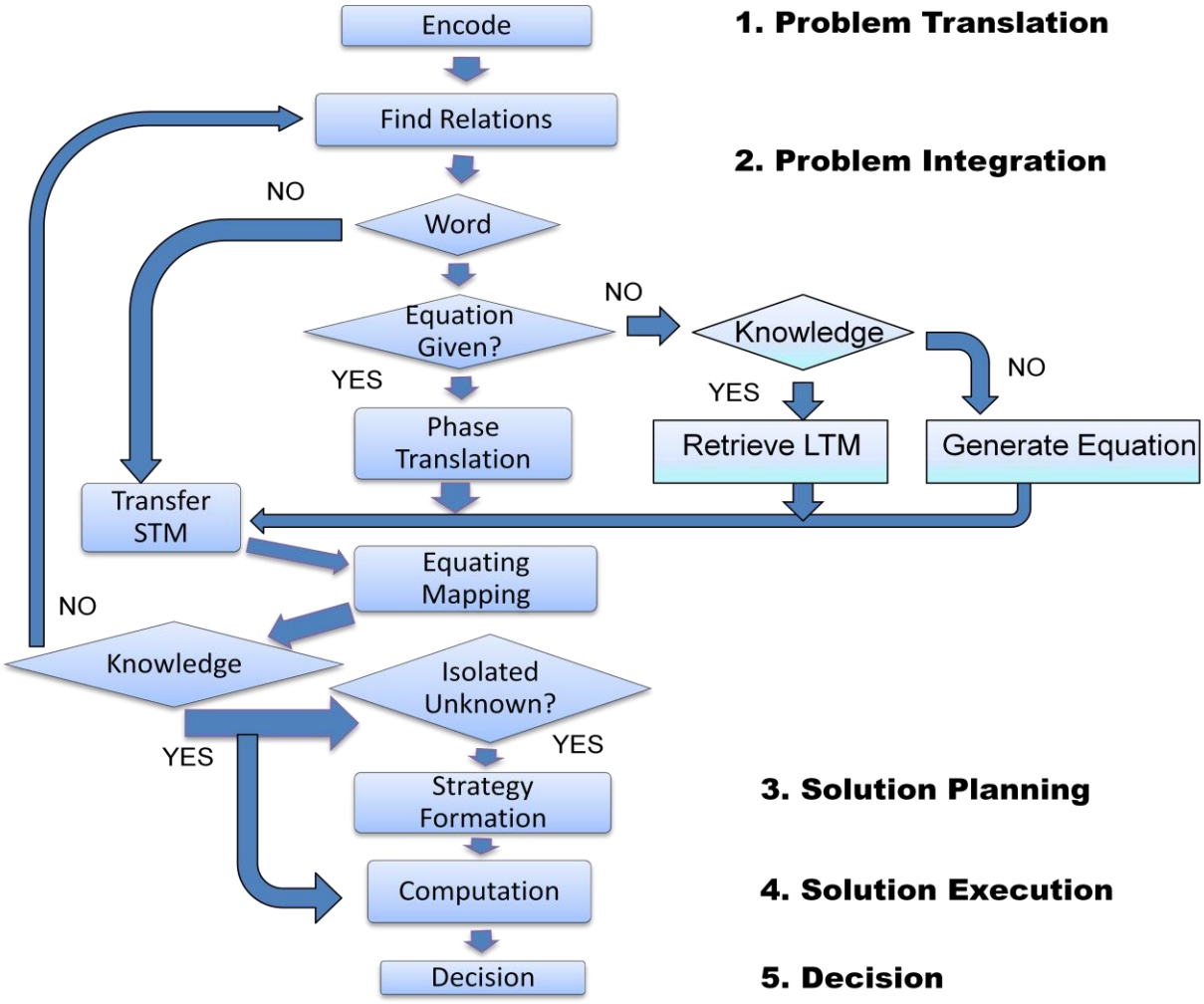


Figure 2.1 Cognitive Model of Mathematical Problem Solving

Cognitive classification of items based on the depth of knowledge levels has become a main feature of standards-based assessment of mathematical achievement. However, because it has not yet been empirically proven that items with greater complexity based on the depth of knowledge levels are more difficult, getting reliable and

valid item classifications is still challenging (Embretson & Daniel, 2008). Embretson (2006) developed several variables to represent cognitive complexity in mathematical problem solving as following five processing stages based on Mayer, Larkin, and Kadane's (1984) theory: (1) problem translation, (2) problem integration, (3) solution planning, (4) solution execution, and (5) decision. Figure 2.1 presents a flow diagram of the five processing stages of mathematical problem solving.

In the first step, "problem translation stage," a single variable, encoding, was scored as the sum of the number of words, terms and operations in the stem. In the second step, "problem integration stage," several variables were scored to represent the processing difficulty for items in which the equation was not directly given. These variables include (a) translating equations from words, (b) the number of knowledge principles or equations to be recalled, (c) the maximum grade level of knowledge principles to be recalled, (d) generating unique equations or representations for the problem, and (e) visualization of relationships (when a diagram is not provided, thus requiring visualization). If the requisite equation is given directly in a question, then the Problem Integration stage, which generates the equation, is not necessary. In the third step, "solution planning stage," two variables, the number of subgoals required to get the final solution and the relative definition of unknowns determine item difficulty. The number of subgoals depends on how many additional variables must be evaluated prior to solving the main task. Relative definition is scored when the equations define the variables only relatively compared to other values in the question. In the fourth step, "solution execution stage," item difficulty can be predicted by the level of procedural knowledge and the number of computations needed for a correct solution of the goal and

all subgoals. In the last step, "decision stage," extra confirmation processing can predict item difficulty when relationships between a solution and the provided answers require extensive consideration.

In conclusion, the cognitive model of mathematical problem solving described above has been empirically supported by many studies (e.g., Embretson, 2003; Embretson & Daniel, 2008; Mayer et al., 1984). It appears sufficiently applicable to a broad bank of mathematical problems. Therefore, the frameworks of "depth of knowledge" and "cognitive model of mathematical problem solving" are very helpful for understanding the item difficulty of mathematical achievement tests.

2.2 Cognition Development for Mathematical Achievement

Because item difficulty primarily depends on the cognitive complexity, development perspectives on children's cognitive will be reviewed. It will be an important background to understand students' cognition development for mathematical achievement.

2.2.1 Developmental Perspectives on Cognition

Among the various theories about human learning (i.e., behaviorism, social cognitive theory, cognitivism, information processing, constructivism, contextualism, and humanism), development perspectives on children's cognition are important background to understand cognition development and children's mathematical achievement. A cognitive development theorist and pioneer in individual constructivism, Jean Piaget, proposed that cognitive development proceeds through four stages which are sensorimotor stage (birth until age 2), preoperational stage (age 2 until age 6 or 7),

concrete operational stage (age 6 or 7 until age 11 or 12), and formal operational stage (age 11 or 12 through adulthood). He insisted that through interacting with and reflecting on their physical and social words, children self-construct increasingly complex understandings and reasoning abilities with age (Piaget & Inhelder, 1969). At about same time, Russian psychologist Lev Vygotsky, a developer of socio-cultural theory, suggested that society and culture give a wide variety of concepts, strategies, and other cognitive "tools" which children gradually begin to use in thinking about and dealing with everyday tasks and problems. Vygotsky placed much of the foundation for a contextual view that has come to be known as the socio-cultural viewpoint. Piaget's and Vygotsky's view of cognitive development share several common themes, yet they also have important theoretical differences.

The common themes between Piaget and Vygotsky include qualitative changes in the nature of thought, challenge, readiness, and the importance of social interaction. First, in the view of qualitative changes in the nature of thought, Piaget and Vygotsky pointed out that children gain more complex reasoning processes over time, in other words, children think differently at different ages. Second, in the view of challenge, Piaget thought that children develop more sophisticated knowledge and thought process only when they meet phenomena they cannot adequately understand using their existing schemes, while Vygotsky's concept was that children can be improved when they have the assistance of the more advanced and competent people than themselves. Third, Piaget and Voygotsky thought that any child would be cognitively ready for some experiences but not ready for others. Piaget insisted that children cannot learn from an experience until they have begun the transition into a stage that allows them to deal with and

conceptualize that experience appropriately. Also, Vygotsky proposed that there are limits on the tasks that children can reasonably hold at any particular time. Fourth, for the importance of social interaction, Piaget explained that the people in a child's life can present information and arguments that make disequilibrium and foster greater perspective taking. Also, Vygotsky proposed that social interactions give the very groundwork for thought processes. According to Vygotsky, children internalize the processes they use when they talk with other persons until they can use them ultimately independently (Ormrod, 2008).

The key theoretical differences between Piaget and Vygotsky's theories of cognitive development are as follows: First, to what extent is language essential for learning and cognitive development? Piaget viewed that much of cognitive development happens independently of language, while Vygotsky claimed that language is very critical for learning and cognitive development. Children's thought processes are internalized versions of social interactions that are mainly verbal in nature. According to Ormrod (2008), Piaget underestimated the importance of language, while Vygotsky overstated the case for language. Second, what kinds of experiences promote learning and development? Piaget thought that self-exploration promotes learning and development, while Vygotsky thought that guided exploration and instruction support learning and development. Third, what kinds of social interactions are most valuable? Piaget emphasized the benefits of interactions with peers, but Vygotsky placed greater importance on interactions with adults and other more advanced individuals. Last, how influential is culture? In Piaget view, the nature of children's logical thinking skills and the progression of these skills over time are largely independent of the specific cultural

context in which children are raised. However, in Vygotsky's view, culture is of paramount importance in determining the specific thinking skills that children acquire (Ormrod, 2008).

In conclusion, Piaget's and Vygotsky's theories are based for development perspectives on cognition and these theories have some ideas in common that continue to appear in more contemporary views of cognitive development. Yet they have important differences that have led modern researchers to search more deeply into the mechanisms through which children's cognitive processes develop.

2.2.2 Cognitive Development and Children's Mathematical Achievement

Next question would be how the cognitive development is related with children's mathematical achievement. In the Piaget's stage (as shown in Table 2.1), it can be assumed that 7th and 8th graders are in formal operations stage (age 11 or 12 through adulthood) because 7th grade age is normally 12 through 13 years old and 8th grade age is normally 13 through 14 years old. When they enter formal operations stage, they are able to think and reason about things that have little or no basis in physical reality such as abstract concepts, hypothetical ideas, and contrary-to-fact statements. In the mathematical and scientific reasoning abilities, proportional thinking is developed and children begin to understand proportions in the form of fractions, decimals, ratios, and so on (Ormrod, 2008).

However, it may be assumed that 6th grade students are in either concrete operational (age 6 or 7 until age 11 or 12) stage or formal operational stage (age 11 or 12 through adulthood) because normally 6th grade age is 11 through 12 years old. When

children enter the concrete operations stage, they can have a form of logical operations. However, they typically do not achieve conservation of weight until relatively late in concrete operations stage. Even though children of concrete operations stage have advancements in reasoning, they have limitation for applying their logical operations. They only apply to concrete observable objects and events and have difficulty dealing with abstract concepts (Ormrod, 2008).

Table 2.1 *Piaget's Stages of Cognitive Development (Ormrod, 2008)*

	Stage	Age Range	General Description
1	Sensorimotor Stage	Birth until about 2 years old	Schemes primarily entail perceptions and behaviors. Children's understandings of the world are based largely on their physical interactions with it
2	Preoperational Stage	2 until about 6 or 7 years old	Many schemes now have a symbolic quality, in that children can think and talk about things beyond their immediate experience. Children begin to reason about events, although not always in ways that are "logical" by adult standards.
3	Concrete Operations Stage	6 or 7 until about 11 or 12 years old	Children acquire cognitive structures that enable them to reason in logical, adultlike ways about concrete, reality-based situations. They also realize that their own perspectives are not necessarily shared by others.
4	Formal Operations Stage	11 or 12 through adulthood	Children can now think logically about abstract, hypothetical, and contrary-to-fact situations. They acquire many capabilities essential for advanced reasoning in mathematics and science.

In Vygotsky's theory, children can accomplish more difficult tasks when they have the assistance of people more advanced and competent than themselves. In other words, students can typically do more difficult things in collaboration with adults than

they can do on their own. For example, children can solve more difficult mathematical problems when their teacher helps them identify critical problem (Ormrod, 2008). Siegler and Ramani (2008) suggested that numerical board games would be especially helpful for improving young children's mathematical understanding. Playing these board games yields large, rapid, and enduring gains in preschoolers' and young elementary school children's numerical understanding. They insisted that the gains are especially large with preschoolers from low-income backgrounds and this can make to reduce the gap in numerical knowledge that separates less and more affluent children when they begin school.

Unlike Piaget, Neo-Piagetian theorists have combined some of Piaget's ideas with concepts from information processing theory to construct how children's learning and reasoning capabilities change over time (Case, 1985). According to Ormrod (2008), information processing theory is focused on how persons think about the information they receive from the environment, how they perceive the stimuli around them, how they put what they have perceived into their memories, and how they find what they have learned when they need to use it. In other words, information processing theory is similar to how computers process information. Robbie Case, one of Neo-Piagetian theorists, was a highly productive researcher and his central conceptual structure is well regarded. His theory integrated network of concepts and cognitive processes that form the basis for much of children's thinking, reasoning, and learning in particular areas. Over time, these structures undergo several major transformations, each of which marks a child's entry to the next higher stage of development. He considered the nature of children's central conceptual structures with respect to several domains, including number, spatial

relationships, and social thought (Case & Okamoto, 1996). This structure shows an integrated understanding of how such mathematical concepts and operations as numbers, counting, addition, and subtraction are interconnected. Case explained that children's development of central conceptual structure for number is until 10 years old, but their understanding of numbers goes on to develop until adolescence. For example, if teenagers don't know what a half of a third is, it means that they have incomplete conceptual understanding of division and the results (e.g., fractions).

To sum up, there are three kinds of perspectives about cognitive development and children's mathematical achievement. First, in the Piaget's perspective, it is assumed that there will be the different stages between 6^{th} grade and the next two grades (7^{th} and 8^{th}). Because it may be assumed that 7^{th} and 8^{th} graders can be in formal operational stage, while 6^{th} grade students may be in either concrete operational stage or formal operational stage. Second, in the aspect of social-cultural theory, children's mathematical achievement can be improved when they have the assistance of people more advanced and competent than themselves. Therefore, an improvement does not occur in specific grade levels, but a significant change can be found in any of 6^{th}, 7^{th}, or 8^{th} grade. Third, in the aspect of Neo-Piagetian theory, the central conceptual structure of numbers develop only until 10 years old, while the understanding of numbers goes on to develop until adolescence. This theory is limited only to numbers thus, it will not be considered in this study.

CHAPTER 3

FUNDAMENTAL ITEM RESPONSE THEORY (IRT) MODELS

The psychometric models for analyzing test item difficulty and measuring the ability change in this study have been developed based on fundamental item response theory (IRT) models. Therefore, the fundamental IRT models will be reviewed in this chapter.

3.1 Assumptions of IRT Models

Since most psychometric models have been developed based on fundamental item response theory (IRT) models, the overview of fundamental IRT models is needed. Currently, in psychometrics, test theories can be divided into two categories: classical test theory (CTT) and item response theory (IRT). Item difficulty statistics, such as *p*-values in CTT or *b*-values in IRT, are modeled from item difficulty factors. IRT is one of the best-known examples of a statistical modeling approach in psychometrics and educational measurement (Tatsuoka, Corter, & Tatsuoka, 2004). IRT has the desirable features of an alternative test theory. If a given IRT model fits the test data of interest, ability estimates obtained from different sets of items will be comparable. Also, IRT provides the possibility of discerning the strength and weakness of each item in a test while CTT analyzes a scale at the test level. IRT is able to distinguish good and bad items in terms of how accurately an item can measure an examinee's trait at the different trait levels (i.e., *item information*; Embretson & Reise, 2000). Concerning IRT, the most

common item parameters are item difficulty, item discrimination, and a pseudo-guessing parameter.

There are two basic assumptions of IRT models about the data to which the models are applied: appropriate dimensionality and local independence. The first assumption, appropriate dimensionality, means that the number of latent traits measured by the items corresponds to the number of trait parameters in the IRT model. For example, if test items depend on two or more latent traits, then IRT models with a single person trait parameter will not be appropriate. Factor analysis, among other methods, can be used to test the assumption. Models which assume the measurement of more than one trait for examinees' test scores are referred to as multidimensional models (Hambleton, Swaminathan, & Rogers, 1991). Several multidimensional IRT (MIRT) models allow for more than one trait (θ) to be estimated, even though the most widely applied IRT models assume a unidimensional construct for which one θ estimate is sufficient to explain item responses (Reckase, 1997). The unidimensionality assumption is closely related to the second assumption, local independence. The local independence assumption means that when the abilities to influence test scores are controlled, examinees' responses to any of the items are statistically independent. Instead of local independence, conditional independence can be considered. That is, within a given trait level, the probability of getting one item correct is independent of the probability of getting other items correct.

3.2 Fundamentals of IRT Models

IRT models can be classified into two basic categories depending on how the items to be analyzed are scored: binary IRT models and polytomous IRT models. Binary

IRT models are used for analyzing items with dichotomously scored responses (e.g., yes/no or, right/wrong); polytomous models can treat multiple category formats, such as rating scales. IRT models also can be classified into two categories depending on the number of latent trait levels (parameters) to be measured in the models: unidimensional and multidimensional models. A unidimensional IRT model is appropriate for data in which a single common factor underlies item response. Thus, a single latent trait is considered to be sufficient to characterize person differences in unidimensional IRT models, while two or more trait levels represent person differences in unidimensional IRT models.

IRT models were originally developed to handle binary response data with a unidimensional construct. Such binary and unidimensional IRT models have been foundations of the development of polytomous and multidimensional IRT models. The original IRT models include traditional logistic models and traditional normal ogive models. There are three traditional logistic models which are widely applied: the one-parameter logistic (1PL) model or Rasch model, the two-parameter logistic (2PL) model and the three parameter logistic (3PL) model. Traditional normal ogive models are two-parameter normal ogive model and three-parameter normal ogive model.

The Rasch model is based on the logistic distribution, which gives the probability of a response in a simple expression. The model predicts the probability of success for person *s* on item *i* $P(X_{is}=1)$. The assumption of unidimensionality is at the heart of the Rasch model. Thus, in Rasch model, trait level indicates the difficulty level at which the individual is as likely to pass as to fail an item (Embretson & Reise, 2000). The Rasch model may be written as follows:

$$P(X_{is}=1|\theta_s,\beta_i)=\frac{\exp(\theta_s-\beta_i)}{1+\exp(\theta_s-\beta_i)}, \quad (3.1)$$

where

$P(X_{is}=1)$ = the probability that person *s* passes item *i*,

X_{is} = response of person s to item *i* (0, 1),

θ_s = trait level for person s,

β_i = difficulty of item *i*, and

$(\theta_s-\beta_i)$ = logit; the simple difference of trait level and item difficulty.

Equation 3.1 is the typical form of the Rasch model and the 1PL model may be written with a constant item discrimination value, α, as follows:

$$P(X_{is}=1|\theta_s,\beta_i)=\frac{\exp[\alpha(\theta_s-\beta_i)]}{1+\exp[\alpha(\theta_s-\beta_i)]}, \quad (3.2)$$

where X_{is}, θ_s, and β_i are same as in 1PL model and α = a constant for item discrimination. In Equation 3.2, the constant value for item discrimination is freely estimated (Embretson & Reise, 2000). The 1PL model is identical to the Rasch model if the value of the constant item discrimination is fixed to 1. That is the difference between the Rasch model and the 1PL model.

The 2PL model adds item discrimination parameter to the Rasch model as follows:

$$P(X_{is}=1|\theta_s,\beta_i,\alpha_i)=\frac{\exp[\alpha_i(\theta_s-\beta_i)]}{1+\exp[\alpha_i(\theta_s-\beta_i)]}, \quad (3.3)$$

where X_{is}, θ_s, and β_i are same as in 1PL model and α_i = discrimination for item *i*.

The 3PL model adds a lower-asymptote parameter to accommodate guessing possibility as follows:

$$P(X_{is}=1|\theta_s,\beta_i,\alpha_i,\gamma_i)=\gamma_i+(1-\gamma_i)\frac{\exp[\alpha_i(\theta_s-\beta_i)]}{1+\exp[\alpha_i(\theta_s-\beta_i)]}, \quad (3.4)$$

where X_{is}, θ_s, β_i, and α_i are defined as above and γ_i = lower-asymptote (guessing) for item *i*. For example, when an item can be solved by guessing, as in multiple-choice items, the probability of success is substantially greater than zero, even for low trait levels.

There are three normal ogive models: one-, two-, and three-parameter models. As for logistic models, normal ogive models are named for the number of item parameters that they have. Normal ogive models contain the same parameters as their corresponding logistic models; however, the item characteristic curve is produced by a different function. The function is more difficult than the logistic model because the probability of success is given by the cumulative proportion of cases in the normal distribution. The one-parameter normal ogive model can be expressed as:

$$\mathrm{P}(X_{is}=1)=\int_{-\infty}^{Z_{is}}\frac{1}{(2\pi)^{1/2}}\exp\left(-t^2/2\right)dt, \quad (3.5)$$

where

$\int_{-\infty}^{Z_{is}} dt$ = integral notation for the area in the distribution from $-\infty$ to Z_{is},

π = the constant 3.14.

The one-parameter normal ogive model does not have the same theoretical appeal or practical application as both the two-parameter normal ogive model and the three-parameter normal ogive model which often have been applied.

The two-parameter normal ogive model contains the same parameter structure as the 2PL model. That is, Z_{is} is given as follows:

$$Z_{is} = \alpha_1(\theta_s - \beta_i). \quad (3.6)$$

The probability of getting an item (i) correct is a function of ability (θ), the item difficulty parameter (b_i), and item discrimination parameter (α_i). The mathematical function of the two-parameter normal ogive model is:

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i)) = \int_{-\infty}^{\alpha i(\theta s - \beta i)} \frac{1}{(2\pi)^{1/2}} \exp(-t^2/2)dt. \quad (3.7)$$

A lower asymptote may be added to the two-parameter normal ogive model to contain item response data with guessing. The mathematical expression for the three-parameter normal ogive model is:

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i, \gamma_i)) = \gamma_i + (1-\gamma_i)\int_{-\infty}^{\alpha i(\theta s - \beta i)} \frac{1}{(2\pi)^{1/2}} \exp(-t^2/2)dt. \quad (3.8)$$

The polytomous IRT models were introduced later as generalized forms of the binary IRT models. Since Samejima introduced the first polytomous IRT model, graded response model (GRM), in 1969, several polytomous IRT models have been developed. The polytomous IRT models can be divided into two types: the indirect (or difference) models and the direct (or divided-by-total) models. The indirect (difference) models include the GRM and the modified graded response model (M-GRM; Muraki, 1990). The direct (divided-by-total) models are the partial credit model (PCM; Masters, 1982), the generalized partial credit model (G-PCM; Muraki, 1992), the rating scale model (RSM; Andrich, 1978), and the nominal response model (NRM; Bock, 1972). Of these models, the PCM has been widely used and applied to several extensions (e.g., G-PCM, LPCM). The PCM was developed for analyzing test items that need multiple steps and for which it was important to assign partial credit for completing several steps in the solution process. Therefore, the PCM naturally explains item response for achievement

tests (e.g., mathematical problems) where partially-correct answers are possible (Embretson & Reise, 2000). The PCM can be considered as an extension of the 1PL model since it has all the standard Rasch model features. For $x=j$ the category response curves for the PCM can be written as:

$$P(X_{ihs}=1)=\frac{\exp(h\theta_s+\beta_{ih})}{\sum_{h=0}^{H_i}\exp(h\theta_s+\beta_{ih})}, \tag{3.9}$$

where

i = 1,…, I items,

$h = 0, \ldots, H_i$ response categories of item i,

$P(X_{ihs}=1)$ = the probability that person s chooses category h on item i

θ_s = the position of person s on the underlying latent trait, and

β_{ih} = the easiness parameter of category h of item i

β_{i0} and $\sum_{i=1}^{I}\sum_{h=1}^{H_i}\beta_{ih}$ are set to zero for model identification and normalization.

In this study, IRTPRO (item response theory for patient-reported outcomes; Thissen, 2010) was used. IRTPRO is one of the most recent applications for item calibration and test scoring using IRT. This program is designed for the efficient analysis of binary items, including multiple choice or short-answer items scored right, wrong, omitted, or not-presented. Also, it is capable of large-scale production applications with ultimately numbers of items or respondents. In this program, following IRT models are available to apply (Thissen, 2010):

(1) Two parameter logistic (2PL) model (Birnbaum, 1968) [with which equality constraints includes the one-parameter logistic (1PL) (Thissen, 1982)]

(2) Three parameter logistic (3PL) model (Birnbaum, 1968)

(3) Graded response model (GRM) (Samejima, 1969; 1997)

(4) Generalized partial credit model (G-PCM) (Muraki, 1992, 1997)

(5) Nominal response model (NRM) (Bock, 1972, 1997; Thissen, Cai, & Bock, in press)

CHAPTER 4

PSYCHOMETRIC MODELS FOR ITEM DIFFICULTY AND ABILITY CHANGE

There are two categories of psychometric models which are possibly relevant to understanding the levels and sources of mathematical item difficulty and measuring ability change. These models are divided into two categories in this review: (1) structured models for item parameters and (2) longitudinal models.

4.1 Structured Models for Item Parameters

Structured models for item parameters are extensions or generalizations of unidimensional IRT models (Rasch, 2PL, and 3PL IRT models). Fischer (1973) proposed the linear logistic latent trait model (LLTM) to incorporate item content into the prediction of item success. The LLTM is an extension to the unidimensional Rasch (1960) model. Enbretson (1999) proposed the 2PL-Constrained model which contains parameters to represent the impact of stimulus features on item discrimination as well as on item difficulty. Glas and van der Linden (2003) proposed a hierarchical IRT model for item structure. That model is a two-level IRT model to incorporate the validity of within-structure item parameters by assuming that the within-structure items are randomly sampled from the structure. Fischer and Ponocny (1994) developed the linear partial credit model (LPCM) to assess the effects of treatments based on polytomous ordered-response items, while the LLTM is applicable only to binary response data.

4.1.1 Linear Logistic Latent Trait Model (LLTM)

Based on the Rasch model, the linear logistic latent trait model (LLTM; Fischer, 1973) was made to integrate item content into the prediction of item success.

Because the LLTM is a generalization of the Rasch model, the 1PL model is important to understand the LLTM. The LLTM is a unidimensional model in which components are identified from item scores on complexity factors that are postulated to determine item difficulty. The LLTM combines a mathematical model for task components with a latent trait model and can incorporate item difficulty into the prediction of item success. The LLTM can be explained by considering three equations.

First, LLTM is a generalization of the Rasch model (Equation 3.1). Equation 3.1 is the Rash latent trait model and it presents the latent trait model for individual differences. Second, Equation 4.1 is the mathematical model for the task processes. β_i is complexity factors and this linear model predicts item difficulty as follows:

$$\beta_i = \sum_{k=0}^{K} \eta_k q_{ik} = \eta_0 q_{i0} + \eta_1 q_{i1} + \eta_2 q_{i2} + \ldots + \eta_k q_{ik}, \qquad (4.1)$$

where

η_k = the effect of stimulus feature *k*,

q_{ik} = the score (e.g., 0 = absence/ 1 = presence) of stimulus feature *k* of item *i*, and

$\eta_0 q_0$ = the intercept of the equation.

In Equation 4.1, item difficulty is scaled as a location on the latent ability continuum or a trait that determines responses.

Third, Equation 4.2 is the LLTM by combining Equation 3.1 and Equation 4.1. If appropriate content factors can be specified for each item, then parameters to reflect the impact on item difficulty can be estimated directly, as follows:

$$P(X_{is}=1)=\frac{\exp(\theta_s-\sum_{k=0}^{K}\eta_k q_{ik})}{1+\exp(\theta_s-\sum_{k=0}^{K}\eta_k q_{ik})}, \qquad (4.2)$$

where

$P(X_{is}=1)$ = the function of $\theta_s-\sum_{k=0}^{K}\eta_k q_{ik}$.

q_{ik} = the linear components of item stimulus features

η_k = their fixed effects, and

θ_s = the person predictor.

If the number of complexity factors is the same as the number of items, and each item has only one complexity factor, then LLTM is equal to the Rasch latent trait model. The LLTM is a linearly-constrained model of item difficulty because item difficulty is modeled by a smaller number of factors.

For example, assume that five factors such as vocabulary level, syntactic complexity, and the density of three basic types of propositions influence the paragraph comprehension items. The LLTM can be applied to estimate the weights of each factor in item difficulty, if the contribution of each factor can be specified numerically for each item. Therefore, each underlying stimulus factor impacts on the difficulty of an item (Embretson & Reise, 2000).

In conclusion, the LLTM belongs to the Rasch family of IRT models, but item difficulty is replaced with a model of item difficulty. Parameters for item difficulty do not appear in the LLTM, but item difficulty is predicted from a weighted combination of

stimulus features that represent the cognitive complexity of the item. In the LLTM, equal item discrimination parameter (α) and no guessing parameter are assumed as in the Rasch model.

4.1.2 2PL-Constrained Model

The 2PL-Constrained model (Embretson, 1999) includes parameters to represent the impact of stimulus features on item discrimination as well as on item difficulty. According to Embretson (1999), item difficulty and item discrimination were both predictable from item stimulus features. This model is a generalization of the 2PL model, while the LLTM is a generalization of the Rasch IRT model. The 2PL-Constrained model can be explained by considering three equations.

First, the 2PL model (Equation 3.3) is used. Second, Equation 4.3 is the mathematical model for both item discrimination (α_i) and item difficulty (β_i) parameters of the 2PL-model.

$$\alpha_i = \sum_{k=0}^{k} \tau_k q_{ik} \, , \; \beta_i = \sum_{k=0}^{k} \eta_k q_{ik} \, . \tag{4.3}$$

Third, Equation 4.4 is the 2PL-Constrained model by combining Equation 4.3 and Equation 3.3. Item discrimination (α_i) and item difficulty (β_i) parameters of the 2PL-model (Equation 3.3) are replaced with the linear combinations of cognitive variables as follows:

$$P(X_{is} = 1) = \frac{\exp[\sum_{k=0}^{k} \tau_k q_{ik} (\theta_s - \sum_{k=0}^{k} \eta_k q_{ik})]}{1 + \exp[\sum_{k=0}^{k} \tau_k q_{ik} (\theta_s - \sum_{k=0}^{k} \eta_k q_{ik})]}, \tag{4.4}$$

where

$P(X_{is} = 1)$ = the probability that person *s* passes item *i*,

θ_s = the ability of person *s*,

q_{ik} = the score for stimulus feature *k* in item *i*,

τ_k = the weight (or effects) of stimulus feature *k* in item discrimination,

η_k = item difficulty, and

$\tau_0 q_0$ and $\eta_0 q_0$ = the intercepts of the equations.

Compared with the LLTM, the 2PL-Constrained model has the advantage of including design features for item discrimination. An item with higher discriminating power provides, both more information about latent trait (θ) and less measurement error because the changes in trait level have a greater impact on $P(X_{is} = 1)$ on that item. In the 2PL-Constrained model, the item stimulus features affecting the discriminating power and the difficulty of the item can be identified. In brief, the 2PL-Constrained model has cognitive complexity models for both item difficulty and item discrimination. However, the 2PL-Constrained model does have some limitations. This model has no error term in the linear combination of cognitive variables (Equation 4.3), thus unable to include random effect on item discrimination and difficulty parameters.

4.1.3 Hierarchical IRT Model

The hierarchical model (Glass & van der Linden, 2003) with the IRT has led to psychometric models for item response datasets that possess a hierarchical structure. The hierarchical IRT model is based on the 3PL model, except that the parameters represent a common value for a family of items rather than unique parameters for each item.
The probability is given for person *j* passing item *i* from family p, and the item parameters are given for the item family can be written as:

$$P(X_{ij_p} = 1) = c_{ip} + (1 - c_{ip}) \frac{\exp(a_{ip}(\theta_j - b_{ip}))}{1 + \exp(a_{ip}(\theta_j - b_{ip}))}, \quad (4.5)$$

where

a_{ip} = item slope or discrimination of item family *p*,

b_{ip} = the item difficulty of item family *p*,

c_{ip} = lower-asymptote of item family *p*, and

θ_j = ability for person *j*.

Items within family *p* are understood to include the same underlying sources of item difficulty, but differing surface features. For example, in mathematical word problems, varying the exact numbers, objects and so forth, are construct-irrelevant surface features. Substituting surface features can create variability within a family. Thus, the hierarchical model includes estimates for error distributions. Briefly, this model does not apply to item families in which the variants are designed to vary in construct-relevant features that impact item difficulty.

4.1.4 Linear Partial Credit Model (LPCM)

After developing the LLTM, Fischer and Ponocny (1994) developed the linear partial credit model (LPCM) to assess the effects of treatments based on polytomous ordered-response items. The LPCM is designed for analyzing polytomous ordered responses which are commonly used for attitude or self-rating items, while the LLTM which is applicable only to binary response data (e.g., right or wrong). The LPCM can be explained by considering three equations.

First, Equation 3.9 is the partial-credit model (PCM) and the LPCM is based on the PCM which assigns one independent parameter β_{ih} for each response category of an item. Second, β_{ih} is replaced with a linear function of cognitive variables, resulting in:

$$\beta_{ih} = \sum_{k=0}^{k} \alpha_k w_{ihk} \,, \tag{4.6}$$

where

α_k = the basic parameter which measures the effect of a certain cognitive variable involved or of an experimental treatment on the response,

w_{ihk} = the value (or, dosage of treatment) of the α_k, and

$\alpha_0 w_{ih0}$ = a normalization constant of the equation.

Notice that α_k and w_{ihk} in the LPCM correspond to η_k and q_{ik} in the LLTM (Equation 4.2), respectively.

Third, the combination of Equation 4.6 with Equation 3.9 is the LPCM.

$$P(X_{ihs} = 1) = \frac{\exp(h\theta_s + \sum_{k=0}^{k} \alpha_k w_{ihk})}{\sum_{h=0}^{H_i} \exp(h\theta_s + \sum_{k=0}^{k} \alpha_k w_{ihk})}. \tag{4.7}$$

Like the LLTM, the LPCM assumes equal α- parameter and no guessing parameter because the LPCM is the extension of the PCM that can be considered as an extension of the Rasch model. One of the nice features of the LPCM is its great flexibility to permit the w_{ihk} parameter for each response category, thus it results in a good fit for complex data. However, the LPCM does have some limitations for application. This model does not fit the simplicity and sometimes it is very hard for interpretation because there is constantly a trade-off between the flexibility and the simplicity of a model.

4.2 Longitudinal Models

Wilson (1985) suggested the SALTUS model for developmental or mastery data. Like other IRT models, SALTUS predicts gradual increases in item success with increasing trait level. A mixed population Rasch model (MIRA) was proposed by Rost (1990) and MIRA combines IRT with latent class analysis. The Multidimensional Rasch Model for Learning and Change (MRMLC) was proposed by Embretson (1991) Additionally, Embretson (1997) proposed structured latent trait models (SLTM).

However, the probabilities increase drastically for some types of items when a person reaches a certain stage. To predict item responses, these models include both trait and class parameters. Even though these models seize only one trait level for each individual, they are classified as multidimensional because of for following reasons. First, each individual's response probability is predicted with not only trait level and item parameters, but also class membership parameters. Second reason is that item difficulty orders are different from individuals like multidimensional models (Embretson & Reise, 2000).

4.2.1 SALTUS Model

Wilson (1985) proposed the SALTUS model (*saltus* means "to leap" in Latin) for developmental or mastery data. In addition, he formulated developmental stage parameters in a Rasch model. Because reaching a certain stage implied a sudden transition in success on an entire class of items, a traditional IRT model frequently does not fit these data. People differ in their acquirement of rules so that when certain rules are

mastered, they solve specific items quickly. However, other items stay unaffected because they do not engage those rules (Embretson & Reises, 2000).

In the SALTUS, the probabilities for success on items gradually increase as the trait levels increase as in other IRT models. However, the probabilities increases significantly when an individual get to a certain stage for some types of items. To model the better achievement for items that are influenced by a particular stage, a parameter is added as follows:

$$P(X_{is} = 1) = \prod_m \frac{\exp(\theta_s - \beta_i + \zeta_{h(s)k(i)})}{1 + \exp(\theta_s - \beta_i + \zeta_{h(s)k(i)})}, \quad (2.20)$$

where

$\zeta_{h(s)k(i)}$ = increased success for item type *k* in stage *h*.

The item characteristic curve (ICC) depends on the developmental stage is the main implication of this model. What make a distinction for the groups is a different pattern of item difficulties or endorsements.

4.2.2 Mixed Population Rasch Model

A mixed population Rasch model (MIRA) was proposed by Rost (1990); MIRA joins IRT with latent class analysis. Latent class model means the model which gives a mastery pattern as the individual's skill profile, while latent trait model refers to the model which places individuals on a continuous scale for each skill including unidimensional and multidimensional IRT models. According to Embretson and Reise (2000), the latent classes can scale and order item difficulties within a different way.

According to a proportion, γ_h , for each latent class, the classes are measured as mixed in the observed sample. The sum of proportions equals one and model is following as:

$$P(X_{is} = 1) = \sum_h \gamma_h \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)} \ . \quad (4.12)$$

In conclusion, MIRA identifies latent classes, each with distinct item difficulty patterns that are required to fit item response data. The classes are mixed in the observed sample. MIRA parameters contain class proportions and class-specific item difficulties make to maximize the likelihood of the item response data.

4.2.3 Multidimensional Rasch Model for Learning and Change (MRMLC) & MRMLC+

Multidimensional Rasch Model for Learning and Change (MRMLC) was proposed by Embretson (1991). This model is an appropriate model for repeated measurement data in which both the means and standard deviations are increasing over occasions. Also, the data must have a simplex correlation structure. MRMLC is appropriate for data on learning potential assessment or longitudinal studies of ability. The MRMLC can be written as

$$P(X_{i(k)j} = 1) = \frac{\exp(\sum_{m=1}^{k} \theta_{jm} - b_i)}{1 + \exp(\sum_{m=1}^{k} \theta_{jm} - b_i)} \ , \quad (4.13)$$

where

$X_{i(k)j}$ = the response of person *j* to item *i* when presented under condition *k*,

θ_{jm} = ability for person *j* on ability *m*, which is collected into a vector θ_j , and

b_i = item difficulty of item i.

Embretson postulated the involvement of M abilities in item responses within K occasions. Specifically, the MRMLC is based on the assumptions: (a) on the first occasion ($k = 1$) only an initial ability is involved in the item responses and (b) on later occasions ($k > 1$), ability plus $k - 1$ additional abilities are involved in the performance. Thus, the number of abilities increases at each time point (occasion). MRMLC+ (Embretson, 1995) is an extension of MRMLC to include the structural model of item difficulty. As in LLTM, item difficulty is predicted from a weighted combination of stimulus features that represent the cognitive complexity of the item. MRMLC+ can be expressed as follows:

$$P(X_{i(k)j} = 1) = \frac{\exp(\sum_{m=1}^{k} \theta_{jm} - \sum_{n} s_{in}\phi_n + f)}{1 + \exp(\sum_{m=1}^{k} \theta_{jm} - \sum_{n} s_{in}\phi_n + f)}, \quad (4.14)$$

where

s_{in} = value on the n stimulus feature of item i,

ϕ_n = impact (weight) of item stimulus factor n on item difficulty, and

f = normalization constant.

Thus, MRMLC+ contains a structural model for item difficulty in addition to the structural model for person ability in MRMLC.

4.2.4 Structured Latent Trait Models (SLTM)

Embretson (1997) proposed structured latent trait models (SLTM) because the traditional Rasch model frequently did not fit the data in ways that would propose

processing differences. According to her, one individual's changes in memory load strongly influenced their item-solving probabilities, while another individual's change in memory load might little change their item-solving probabilities.

A general SLTM may be given as follows:

$$P(X_{i(k)j} = 1) = \frac{\exp(\sum_m \lambda_{i(k)m}\theta_{jm} - \sum_k b_{ik})}{1 + \exp\sum_m \lambda_{i(k)m}\theta_{jm} - \sum_k b_{ik})}, \quad (4.15)$$

where

$X_{i(k)j}$ = the response of person *j* to item *i* when presented under condition *k*,

θ_{jm} = ability for person *j* on ability *m*, which is collected into a vector θ_j,

b_{ik} = difficulty contribution of condition *k* on item i, and

$\lambda_{i(k)m}$ = the weight of ability *m* in item *i* under condition *k*, which is collected into a weight matrix Λ.

An important feature of the SLTM is that the involvement of a particular processing ability, θ_{jm}, depends on the item $\lambda_{i(k)m}$. If the item discriminations are constrained to the same value within each combination of condition *k* and ability *m*, then the resulting SLTM is a Rasch-family model.

The SLTM includes several sub-models. SLTM-1 is one of the Rasch-family model and item difficulties (b_{ik}) are constrained over all abilities. SLTM-2 also belongs to the Rasch-family models and its structures have fixed values (i.e., constants other than 0 or 1 are allowed). This values stand for structured comparisons of performance across occasions. SLTM-3 is also general model structures and item difficulties are constrained across abilities. SLTM-3 is helpful for data in which items are not equally discriminating

on the latent trait. SLTM-4 is also the same general models as SLTM-1and belongs to the Rasch-family model. However, item difficulties can be changed across occasions. Thus, SLTM-4 is practical for conditions that contain person changes that are not equally important on all items (Embretsonn, 1997).

4.3 Summary and Evaluation of the Models

The list of the psychometric models reviewed in this chapter is presented for each category in Table 4.1. First, in the structured model category, LLTM, 2PL-Constrained model, and LPCM are appropriate for understanding the levels and sources of item difficulty since they contain the structural model of item difficulty. Item difficulty can be predicted from a weighted combination of item stimulus features, that is the sources of item difficulty, in these models. LLTM, 2PL-Constrained model are unidimensional IRT models, thus being relatively simple and easy to apply to the test design; unidimensional IRT models fit well with most achievement test data.

Second, in the longitudinal model category, MRMLC and MRMLC+ resolved some basic problems in the classical measurement of individual change score such as the reliability paradox and the scaling problem. Additionally, MRMLC+ has a structural model of item difficulty, thus being useful for understanding the levels and sources of mathematical item difficulty. MRMLC+ permits linking individual learning to changes in cognitive processing and knowledge structures. In this category, SLTM is also applicable to understanding item difficulty. SLTM was developed based on the idea that an individual’s change in memory load influenced their item-solving probabilities depending

on occasions. Of the four sub-models of SLTM, SLTM-4 appears practical since item difficulty can be estimated differently across occasions.

Table 4.1 *List of Reviewed Psychometric Models based on Two Categories*

Categories	*Psychometric Models*	*Understanding Item Difficulty*	*Measuring Change*
Structured models	LLTM (Fischer, 1973)	Appropriate	Appropriate
	2PL-Constrained model (Embretson, 1999)	Appropriate	Applicable
	Hierarchical IRT Model (Glass & van der Linden, 2003)		
	LPCM (Fischer & Ponocny, 1994)	Appropriate	Applicable
Longitudinal models	SALTUS Model (Wilson, 1985)		Appropriate
	MIRA (Rost, 1990)		Appropriate
	MRMLC (Embretson, 1991)		Appropriate
	MRMLC+ (Embretson, 1995)	Appropriate	Appropriate
	SLTM (Embretson, 1997)	Applicable	Appropriate

In conclusion, LLTM and its extensions such as 2PL-Constrained model, LPCM, and MRMLC+ are appropriate for understanding item difficulty. SLTM are applicable to understand the levels and sources of mathematical item difficulty. For analyzing ability change, all longitudinal models as well LLTM are appropriate, while 2PL-Constrained model and LPCM are applicable.

CHAPTER 5

STANDARDS, BENCHMARKS, AND INDICATORS OF MATHEMATICAL ACHIEVEMENT

5.1 Identifying Mathematical Standards and Benchmarks

Mathematical achievement standards and benchmarks are very important milestones. Standards for middle school mathematics are needed for all state students to learn mathematical content and skills that are used to solve a variety of problems. As shown in Table 5.1, there are four levels of standards. Standard 1 is "number and computation"; students use numerical and computational concepts and procedures in a variety of situations. Standard 2 is "algebra"; students use algebraic concepts and procedures in a variety of situations. Standard 3 is "geometry"; students use geometric concepts and procedures in a variety of situations. Standard 4 is "data"; students use concepts and procedures of data analysis in a variety of situations. These standards are applied to all 6^{th}, 7^{th}, and 8^{th} grades (Kansas State Department of Education, 2003).

Table 5.1 *Standards and Description for Middle School Mathematics*

Standard	*Standard Description*
1. Number and computation	The student uses numerical and computational concepts and procedures in a variety of situations.
2. Algebra	The student uses algebraic concepts and procedures in a variety of situations.
3. Geometry	The student uses geometric concepts and procedures in a variety of situations.
4. Data	The student uses concepts and procedures of data analysis in a variety of situations.

As shown in Table 5.2, Standard 1(number and computation) includes four benchmarks: Benchmark 1.1 (number sense), Benchmark 1.2 (number system and their properties), Benchmark 1.3 (estimation), and Benchmark 1.4 (computation).

Table 5.2 *Standard 1 (Number and computation)'s Benchmarks and Descriptions*

Benchmark	*Grade*	*Description*
1.1. Number Sense	6th	The student demonstrates number sense for rational numbers and simple algebraic expressions in one variable in a variety of situations.
	7th	The student demonstrates number sense for rational numbers, the irrational number pi, and simple algebraic expressions in one variable in a variety of situations.
	8th	The student demonstrates number sense for real numbers and simple algebraic expressions in one variable in a variety of situations.
1.2. Number system and their properties	6th	N/A in this test
	7th	N/A in this test
	8th	The student demonstrates an understanding of the real number system; recognizes, applies, and explains their properties; and extends these properties to algebraic expressions.
1.3. Estimation	6th	The student uses computational estimation with rational numbers and the irrational number pi in a variety of situations.
	7th	N/A in this test
	8th	N/A in this test
1.4 Computation	6th	The student models, performs, and explains computation with positive rational numbers and integers in a variety of situations.
	7th	The student models, performs, and explains computation with rational numbers, the irrational number pi, and first-degree algebraic expressions in one variable in a variety of situations.
	8th	The student models, performs, and explains computation with rational numbers, the irrational number pi, and algebraic expressions in a variety of situations.

As shown in Table 5.3, Standard 2 (algebra) has four benchmarks: Benchmark 2.1 (patterns), Benchmark 2.2 (variables/equations and inequalities), Benchmark 2.3 (functions), and Benchmark 2.4 (models).

Table 5.3 *Standard 2 (Algebra)'s Benchmarks and Descriptions*

Benchmark	*Grade*	*Description*
2.1. Patterns	6th	The student recognizes, describes, extends, develops, and explains the general rule of a pattern in variety of situations.
	7th	The student recognizes, describes, extends, develops, and explains the general rule of a pattern in variety of situations.
	8th	N/A in this test
2.2. Variables, Equations, and Inequalities	6th	The student uses variables, symbols, positive rational numbers, and algebraic expressions in one variable to solve linear equations and inequalities in a variety of situations.
	7th	The student uses variables, symbols, rational numbers, and simple algebraic expressions in one variable to solve linear equations and inequalities in a variety of situations.
	8th	The student uses variables, symbols, rational numbers, and algebraic expressions to solve linear equations and inequalities in a variety of situations.
2.3. Functions	6th	N/A in this test
	7th	N/A in this test
	8th	The student recognizes, describes, and analyzes constant, linear and nonlinear relationships in a variety of situations.
2.4 Models	6th	N/A in this test
	7th	N/A in this test
	8th	The student generates and uses mathematical models to represent and justify mathematical relationships found in a variety of situations.

As shown in Table 5.4, four benchmarks of Standard 3 (geometry) are Benchmark 3.1 (geometric figures and their properties), Benchmark 3.2 (measurement and estimation), Benchmark 3.3 (transformational geometry), and Benchmark 3.4 (geometry from an algebraic perspective).

Table 5.4 *Standard 3 (Geometry)'s Benchmarks and Descriptions*

Benchmark	*Grade*	*Descriptions*
3.1. Geometric Figures and Their Properties	6th	The student recognizes geometric figures and compares their properties in a variety of situations.
	7th	Same as above
	8th	Same as above
3.2. Measurement and Estimation	6th	The student estimates, measures, and uses measurement formulas in a variety of situation.
	7th	The student estimates, measures, and uses measurement formulas in a variety of situation.
	8th	N/A in this test
3.3. Transformational Geometry	6th	The student recognizes and performs transformations on two- and three-dimensional geometric figures in a variety of situations.
	7th	The student recognizes and performs transformations on two- and three-dimensional geometric figures in a variety of situations.
	8th	N/A in this test
3.4. Geometry from an Algebraic Perspective	6th	The student relates geometric concepts to a number line and a coordinate plane in a variety of situations.
	7th	N/A in this test
	8th	The student uses an algebraic perspective to examine the geometry of two-dimensional figures in a variety of situations.

As shown in Table 5.5, Standard 4 (data) includes two benchmarks: Benchmark 4.1 (probability) and Benchmark 4.2 (statistics).

Table 5.5 *Standard 4 (Data)'s Benchmarks and Descriptions*

Benchmark	*Grade*	*Description*
4.1. Probability	6th	The student applies the concepts of probability to draw conclusions and to make predictions and decisions including the use of concrete objects in a variety of situations.
	7th	N/A in this test
	8th	The student applies the concepts of probability to draw conclusions, generate convincing arguments, and make predictions and decisions including the use of concrete objects in a variety of situations.
4.2. Statistics	6th	N/A in this test
	7th	The student collects, organizes, displays, and explains numerical (rational numbers) and non-numerical data sets in a variety of situations with a special emphasis on measures of central tendency.
	8th	The student collects, organizes, displays, and interprets numerical (rational) and non-numerical data sets in a variety of situations.

5.2 Comparison of Benchmarks and Indicators among 6th, 7th, and 8th Grades

Notice that description of a benchmark is slightly different for each grade level except Benchmark 3.1(geometric figures and their properties) whose description is same for all the grade levels (see Table 5.4). It should be also noted that the fourteen benchmarks are not applied for all 6th, 7th, and 8th grades in this test (see Table 5.6).

A total of four standards and fourteen benchmarks exist for middle school mathematics: Standard 1(number and computation), Standard 2 (algebra), and Standard 3 (geometry) have four benchmarks each while Standard 4 (data) includes two benchmarks.

Table 5.6 *Standards and Benchmarks for Middle School Mathematics*

Standard	***Benchmark***	***6th Grade***	***7th Grade***	***8th Grade***
1. Number and Computation	1.1. Number sense	Yes	Yes	Yes
	1.2.Number systems and their properties	N/A	N/A	Yes
	1.3. Estimation	Yes	N/A	N/A
	1.4. Computation	Yes	Yes	Yes
2.Aalgebra	2.1. Patterns	Yes	Yes	N/A
	2.2. Variables, Equations, and Inequalities	Yes	Yes	Yes
	2.3. Functions	N/A	N/A	Yes
	2.4. Models	N/A	N/A	Yes
3.Geometry	3.1. Geometric Figures and Their Properties	Yes	Yes	Yes
	3.2. Measurement and Estimation	Yes	Yes	N/A
	3.3. Transformational Geometry	Yes	Yes	N/A
	3.4. Geometry from an Algebraic Perspective	Yes	N/A	Yes
4. Data	4.1. Probability	Yes	N/A	Yes
	4.2.Statistics	N/A	Yes	Yes

The number of benchmarks within the standards is different depending on grades as shown in Table 5.6. Also, those benchmarks include different indicators for each grade level. For analyzing the change in mathematical achievement in 6th grade through 8th grade, all 6th, 7th, and 8th grade mathematical achievement results are needed. Thus, in this study, four benchmarks 1.1 (number sense), 1.4 (computation), 2.2 (variables/equations and inequalities), and 3.1 (geometric figures and their properties) will be focused for analyzing the change in mathematical achievement in from 6th to 8th grade.

There are three categories: for only one grade, two grades, and all grades. First, some benchmarks are applied for only one grade. For example, benchmark 1.3 (estimation) is used only for 6th grade. Benchmark 1.2 (number systems and their properties), Benchmark 2.3 (functions), and Benchmark 2.4 (models) are used only for 8th grade. Second, some benchmarks are applied for two grades. Benchmark 2.1 (patterns), Benchmark 3.2 (measurement and estimation), and Benchmark 3.3 (transformational geometry) are for both 6th and 7th grades. Benchmark 3.4 (geometry from an algebraic perspective) and Benchmark 4.1 (probability) are for both 6th and 8th grades. Benchmark 4.2 (statistics) is only for 7th and 8th grades. However, four benchmarks are applied for all grades. Benchmark 1.1 (number sense), Benchmark 1.4 (computation), Benchmark 2.2 (variables/equations and inequalities), and Benchmark 3.1 (geometric figures and their properties) are applied for all 6th, 7th, and 8th grades.

In this study, six benchmarks (1.1 number sense, 1.4 computation, 2.2 variables/equations and inequalities, 3.1 geometric figures and their properties, 3.4 geometry from an algebraic perspective, and 4.1 probability will be focused on for analyzing the change in mathematical achievement in from 6th to 8th grade. Two benchmarks (3.4 geometry from an algebraic perspective and 4.1 probability) are applied for both 6th and 8th grades, while four benchmarks (1.1 number sense, 1.4 computation, 2.2 variables/equations and inequalities, and 3.1 geometric figures and their properties are for all 6th, 7th, and 8th grades. The change in the set of benchmarks becomes more cognitively complex for higher grades. Thus, indicators of a benchmark are also different for each grade level because of change for cognitively complex. Each grade (6th, 7th, and

8th grade)'s indicators of the benchmarks 1.1, 1.4, 2.2, 3.1, 3,4, and 4.1 are presented in Table 5.7, 5.8, 5.9, 5.10, 5.11, and 5.12, respectively.

First, the description of Benchmark 1.1 (number sense) for 6th grade is that the student demonstrates number sense for rational numbers and simple algebraic expressions in one variable in a variety of situations. However, for 7th grade, it is that the student demonstrates number sense for rational numbers, the "irrational number pi", and simple algebraic expressions in one variable in a variety of situations. Also, for 8th grade, student should demonstrate number sense for "real numbers" and simple algebraic expressions in one variable in a variety of situations. For indicators, as shown in Table 5.7, Benchmark 1.1 includes different indicators for each of 6th, 7th, and 8th grades.

Table 5.7 *Indicators of the Benchmark 1.1 of Standard 1*

Benchmark	*Grade*	*Indicator Description*
1.1. Number Sense	6th	1.1. 1. Compares and orders : a) integers; b) fractions greater than or equal to zero; c) decimals greater than or equal to zero through thousandths place 1.1.2. Knows and explains numerical relationships between percents
	7th	1.1.1. Generates and/or solves real-world problems using: a) equivalent representations of rational numbers and simple algebraic expressions: b) addition, subtraction, multiplication, and division of rational numbers with a special emphasis on fractions and expressing answers in simplest form
	8th	1.1.1. Knows and explains what happens to the product or quotient when: a) a positive number is multiplied or divided by a rational number greater than zero and less than one; b) a positive number is multiplied or divided by a rational number greater than one; c) a nonzero real number is multiplied or divided by zero

Second, as shown in Table 5.8, Benchmark 1.4 includes different indicators for each grade level.

Table 5.8 *Indicators of the Benchmark 1.4 of Standard 1*

Benchmark	*Grade*	*Indicator Description*
1.4 Computation	6th	1.4.1. Performs and explains these computational procedures: a) divides which numbers through a 2-digit divisor and a 4-digit dividend and expresses the remainder as a whole number, fraction, or decimal 1.4.2. Generates and/or solves one- and two-step real-world problems with rational numbers using the computational procedures: b) addition, subtraction, multiplication, and division of decimals through hundredths place
	7th	1.4.1. Performs and explains these computational procedures: a) adds and subtracts decimals from ten millions place through hundred thousandths place; b) multiplies and divides a four-digit number by a two-digit number using numbers from thousands place through thousandths place; c) multiplies and divides using numbers from thousands place through thousandths place by 10; 100; 1,000; 0.1; 0.01; 0.001; or signal-digit multiplies of each; d) adds, subtracts, multiplies, and divides fractions and expresses answers in simplest form 1.4.2. Finds percentages of rational numbers
	8th	1.4.1. Performs and explains these computational procedures with rational numbers: a) addition, subtraction, multiplication, and division of integers; b) order of operations (evaluates within grouping symbols, evaluates powers to the second or third power, multiplies or divides in order from left to right, then adds or subtracts in order from left to right) 1.4.2. Generates and/or solves one- and two-step real-world problems using computational procedures and mathematical concepts: a) rational numbers; b) the irrational number pi as an approximation; c) applications of percents

The description of Benchmark 1.4 (computation) for 6th grade is that the student models, performs, and explains computation with positive rational numbers and integers in a variety of situations. For 7th grade, it is more complex than 6th grade. For example, the student models, performs, and explains computation with "rational numbers, the irrational number pi, and first-degree algebraic expressions in one variable" in a variety of situations. For 8th grade, it is more complex than 7th grade. The student should model, perform, and explain computation with "rational numbers, the irrational number pi, and algebraic expressions" in a variety of situations.

Third is Benchmark 2.2 (variables/equations and inequalities) in Table 5.9.

Table 5.9 *Indicators of the Benchmark 2.2 of Standard 2*

Benchmark	*Grade*	*Indicator Description*
2.2. Variables, Equations, and Inequalities	6th	2.2.1. Represents real-world problems using variables and symbols to: b) write and/or solve one-step equations (addition, subtraction, multiplication, and division)
	7th	2.2.1. Knows the mathematical relationship between ratios, proportions, and percents and how to solve for a missing term in a proportion with positive rational number solutions and monomials 2.2.2. Evaluates simple algebraic expressions using positive rational numbers 2.2.3. Represents real-world problems using variables and symbols to write linear expressions, one- or two-step equations
	8th	2.2.1. Solves: a) one- and two-step linear equations in one variable with rational number coefficients and constants intuitively and/or analytically 2.2.2. Represents real-world problems using: a) variables, symbols, expressions, one- or two- step equations with rational number coefficients and constants

The description of Benchmark 2.2 (variables, equations, and inequalities) for 6th grade is that the student uses variables, symbols, "positive rational numbers", and algebraic expressions in one variable to solve linear equations and inequalities in a variety of situations. For 7th grade, it is more complex because the student uses variables, symbols, "rational numbers, and simple algebraic expressions "in one variable to solve linear equations and inequalities in a variety of situations. For 8th grade, it is more complex than 7th grade since the student should use variables, symbols, "rational numbers, and algebraic expressions" to solve linear equations and inequalities in a variety of situations.

Table 5.10 *Indicators of the Benchmark 3.1 of Standard 3*

Benchmark	*Grade*	*Indicator Description*
3.1. Geometric Figures and Their Properties	6th	3.1.1. Classifies: a) angles as right, obtuse, acute, or straight; b) triangles as right, obtuse, acute, scalene, isosceles, or equilateral
	7th	3.1.1. Identifies angle and side properties of triangles and quadrilaterals: a) sum of the interior angles of any triangle is 180˚; b) sum of the interior angles of any quadrilateral is 360˚; c) parallelograms have opposite sides that are parallel and congruent; d) rectangles have angles 90˚ opposite sides are congruent; e) rhombi have all sides the same length, opposite angles are congruent; f) squares have angles of 90˚, all sides congruent; g) trapezoids have one pair of opposite sides parallel and the other pair of opposite sides are not parallel
	8th	3.1.1. Uses the Pythagorean theorem: a) determine if a triangle is a right triangle; b) find a missing side of a right triangle where the lengths of all three sides are whole numbers 3.1.2. Solves real-world problems by: a) using the properties of corresponding parts of similar and congruent figures

Fourth, as shown in Table 5.10, Benchmark 3.1 (geometric figures and their properties) includes different indicators for each grade level. The description of Benchmark 3.1 for all grade levels is same: The student recognizes geometric figures and compares their properties in a variety of situations. However, Benchmark 3.1(geometric figures and their properties) has different indicators for each of 6th, 7th, and 8th grades because of change for cognitively complex.

Fifth, as shown in Table 5.11, Benchmark 3.4 (geometry from an algebraic perspective) includes different indicators for 6th grade and 8th grade level. The description of Benchmark 3.4 (geometry from an algebraic perspective) for 6th grade is that the students relate geometric concepts to a number line and a coordinate plane in a variety of situations. However, for 8th grade, it is that the student uses an algebraic perspective to examine the geometry of two-dimensional figures in a variety of situations.

Table 5.11 *Indicators of the Benchmark 3.4 of Standard 3*

Benchmark	*Grade*	*Indicator Description*
3.4. Geometry from an Algebraic Perspective	6th	3.4.1. Uses all four quadrants of the coordinate plane: a) identify the ordered pairs of integer values on a given graph; b) plot the ordered pairs of integer values
	7th	N/A in this test
	8th	3.4.1. Uses the coordinate plane to: a) list several ordered pairs on the graph of a line and find the slope of the line; b) recognize that ordered pairs that lie on the graph of an equation are solutions to that equation; c) recognize that points that do not lie on the graph of an equation are not solutions to that equation; d) determine the length of a side of a figure drawn on a coordinate plane with vertices having the same x- or y- coordinates

Sixth, as shown in Table 5.12, Benchmark 4.1(probability) also includes different indicators for 6th grade and 8th grade level. The description of Benchmark 4.1 for 6th grade is that the students apply the concepts of probability to draw conclusions and to make predictions and decisions including the use of concrete objects in a variety of situations. However, for 8th grade, it is that the students apply the concepts of probability to draw conclusions, generate convincing arguments, and make predictions and decisions including the use of concrete objects in a variety of situations.

Table 5.12 *Indicators of the Benchmark 4.1 of Standard 4*

Benchmark	*Grade*	*Indicator Description*
4.1. Probability	6th	4.1.1. List all possible outcomes of an experiment or simulation with a compound event composed of two independent events in a clear and organized way 4.1.2. Represents the probability of a simple event in an experiment or simulation using fractions and decimals
	7th	N/A in this test
	8th	4.1.1. Finds the probability of a compound event composed of two independent events in an experiment, simulation, or situation 4.1.2. Makes predictions based on the theoretical probability a) a simple event in an experiment or simulation

In conclusion, some benchmarks are applied for only one grade: Benchmark 1.3 (estimation) is used only for 6th grade while Benchmark 1.2 (number systems and their properties), Benchmark 2.3 (functions), and Benchmark 2.4 (models) are used only for 8th grade. Also, some benchmarks are applied for two grades: Benchmark 2.1 (patterns), Benchmark 3.2 (measurement and estimation), and Benchmark 3.3 (transformational geometry) are for both 6th and 7th grades, Benchmark 3.4 (geometry from an algebraic

perspective) and Benchmark 4.1 (probability) are for both 6th and 8th grades, and Benchmark 4.2 (statistics) is for 7th and 8th grades. However, four benchmarks are applied for all 6th, 7th, and 8th grades. They are Benchmark 1.1 (number sense), Benchmark 1.4 (computation), Benchmark 2.2 (variables/equations and inequalities), and Benchmark 3.1 (geometric figures and their properties). These four benchmarks are very useful for analyzing the shift in the achievement across grades. Benchmark 3.4 (geometry from an algebraic perspective) and Benchmark 4.1 (probability) are also used in this study because they are for both 6th and 8th grades it can be analyzed the shift in the achievement. Therefore, six benchmarks 1.1 (number sense), 1.4 (computation), 2.2 (variables/equations and inequalities), 3.1 (geometric figures and their properties), 3.4 (geometry from an algebraic perspective), and 4.1 (probability) are focused on for analyzing the change in mathematical achievement across grades.

CHAPTER 6

STUDY 1: ANALYSIS OF ITEM DIFFICULTY AND FACTOR STRUCTURE

6.1 Study 1 Purpose and Hypotheses

6.1.1 Study 1 Purpose

Study 1 was conducted as a preliminary study to investigate general properties of the state mathematical achievement tests before measuring students' achievement changes in Study 2. Item difficulties of the mathematical standards and benchmarks as well as their factor structures were carefully examined in this study. Therefore, objectives of this study were (1) to estimate item difficulties of the four standards and the six specific benchmarks (1.1, 1.4, 2.2, 3.1, 3.4, and 4.1), (2) to examine the implications of the four standards for independent factors, using a full information confirmatory factor analysis (FICFA), which also provides θ –estimations in each factor, and (3) to examine the extent to which a general factor underlies all the variables beyond content-specific group factors (i.e., four standards, six specific benchmarks) using item bifactor analysis.

6.1.2 Study 1 Hypotheses

6.1.2.1 Study 1.1 Hypothesis: Item difficulty

Identifying the difficulty of each mathematical standard and benchmark is a fundamental and important step before measuring the mathematical achievement change in Study 2. In this study, item difficulties of the four standards and the six benchmarks were estimated within grades. There are two considerations that lead to hypotheses regarding item difficulties.

First, difficulties of the four mathematical standards (numbers and computation, algebra, geometry, and data) may be different. According to Tatsuoka et al. (2004), U.S. students are weak, especially in geometry. They insisted that success in geometry was found to be highly associated with logical reasoning and other important mathematical thinking skills. Therefore, it is expected that Standard 3 (geometry), in this study, is harder than the other standards. Then, Standard 2 (algebra) may be harder than both Standard 1 (numbers and computation) and Standard 4 (data) because it requires logical or abstract reasoning. However, an alternative view is that difficulties of the four mathematical standards (numbers and computation, algebra, geometry, and data) may be same because the test items are selected to make the standards equally difficult.

Second, it is also expected that item difficulties of the six specific benchmarks are different from each other. Benchmark 3.4 (geometry from an algebraic perspective) may be harder than any of the benchmarks because it requires logical reasoning and some algebraic characteristics. Benchmark 3.1 (geometric figures and their properties) may be harder than the other four benchmarks because the spatial and geometric concepts are required. In addition, Benchmark 2.2 (variables/equations and inequalities) may be harder than the rest of the three benchmarks because of the logic and abstract reasoning requirement. However, similar to the standards, item difficulties of the six specific benchmarks may be same because the test items are selected to make the benchmark equally difficult.

In this study, following two hypotheses were tested based on the view that difficulties of the contents (i.e., standards, benchmarks) may be different.

Hypothesis 1. Difficulties of the four standards are in the following order: Standard 3 (Geometry) > Standard 2 (Algebra) > Standard 4 (Data) > Standard 1 (Number and computation). * harder > easier

Hypothesis 2. Difficulties of the six specific benchmarks are in the following order: Benchmark 3.4 (geometry from an algebraic perspective) > Benchmark 3.1 (geometric figures and their properties) > Benchmark 2.2 (variables/equations and inequalities) > Benchmark 4.1 (probability) > Benchmark 1.4 (computation) > Benchmark 1.1 (number sense). * harder >easier

6.1.2.2 Study 1-2 Hypothesis: Full Information Confirmatory Factor Analysis

Since the mathematical achievement tests in this study were developed based on the four standards as elaborated in Chapter 5, full information confirmatory factor analysis (FICFA) is appropriate to examine the relative independence of the skills that are defined. Thus, exploratory factor analysis (EFA) was not performed. FICFA was carried out to identify the four-factor structure by the four standards for each grade data.

Although each of the mathematical items was originally constructed to involve a specific single standard according to the blueprint, Embretson (2011) found that some items may require the mastery of multiple standards. For example, in reality, some items can involve two standards such as Standard 1 (numbers and computation) and Standard 2 (algebra) or more. A Q-matrix which is incidence matrix of the attributes (e.g., standards, benchmarks) involved in each item was developed by a mathematician who was experienced in state assessment. In this study, the FICFA model without split loadings (non-zero loadings on only one factor) is based on the blueprint while the FICFA model

allowing split loadings (non-zero loadings on more than one factor) is based on the Q matrix. Thus, There are two following hypothesis regarding the FICFA models:

Hypothesis 3. The four factor structure defined by the four standards will be identified for each grade.

Hypothesis 4. The FICFA model with split loadings fits the data better than the FICFA model without split loadings.

6.1.2.3 Study 1-3 Hypothesis: Bifactor Analysis by Content

It is plausible for many types of psychological and educational tests to exhibit a two level-structure of a "general" factor and one or more content-specific group factors (Gibbons, Bock, Hedeker et al., 2007). In this mathematical achievement test, a general factor represents the overall mathematical competency and the content-specific group factors correspond to the specific contents such as the four standards or six benchmarks. The bifactor model constrains each item to have a non-zero loading on the general factor and a secondary loading on no more than one of the group factors. Also, the group factors are assumed uncorrelated with each other (Weiss & Gibbons, 2007). Although the constraint of no split loading in the bifactor model does not fit the multiple-standard involvement of the test items in this study, the item bifactor analysis was used to examine the extent to which a general factor accounts for the data and to examine the extent to which content-specific group factors reflect the four standards or six benchmarks.

Compared with the FICFA, bifactor analysis can provide the information as to whether all the test items commonly represent a general factor (i.e., mathematical competency) beyond the content-specific group factors (four standards or six benchmarks)

and the contribution of each content on the general factor. That is the reason why a bifactor analysis is needed. Existence of the bifactor structure will be examined for each grade. Thus, the following hypothesis is tested:

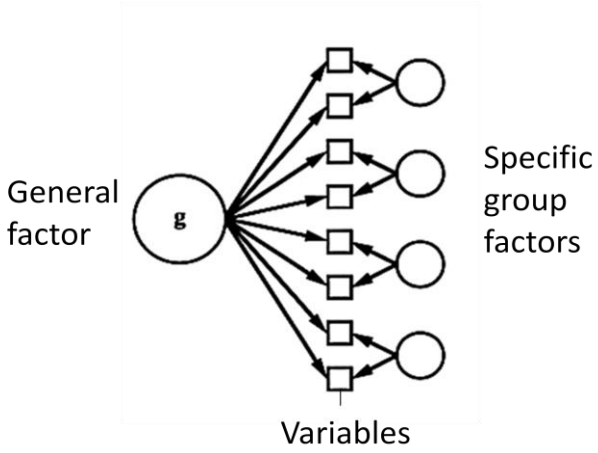


Figure 6.1 Bifactor Model

Hypothesis 5. The bifactor structure which includes a general factor (mathematical competency) and content-specific group factors (four standards or six benchmarks) exists across the grades.

6.2 Study 1 Method

6.2.1 Subjects and Instruments

For longitudinal data, same subjects at different points of time are needed. A random sample of 2667 students in a Midwest state was obtained, who were in sixth, seventh, and eighth grade in 2006, 2007, and 2008, respectively. A mathematical achievement tests were administered online to all the students at the end of each school year.

The mathematical achievement test consists of 86, 84, and 86 items for 6th, 7th, and 8th grade, respectively. All items in the test were multiple-choice items with four alternatives. As shown in Table 6.1, there is assessment framework for middle school mathematics.

Table 6.1 *Assessment Framework for Middle School Mathematics*

Standard	***Benchmark***	***6th Grade***	***7th Grade***	***8th Grade***
Total Objective Questions		86	84	86
1. Number and Computation	1.1. Number sense	10	5	6
	1.2.Number systems and their properties	N/A	N/A	10
	1.3. Estimation	6	N/A	N/A
	1.4. Computation	14	13	12
Number and Computation Subtotal		30	18	28
2.Aalgebra	2.1. Patterns	4	10	N/A
	2.2. Variables, Equations, and Inequalities	8	19	12
	2.3. Functions	N/A	N/A	5
	2.4. Models	N/A	N/A	4
Algebra Subtotal		12	29	21
3.Geometry	3.1. Geometric Figures and Their Properties	6	7	9
	3.2. Measurement and Estimation	14	14	N/A
	3.3. Transformational Geometry	4	4	N/A
	3.4. Geometry from an Algebraic Perspective	6	N/A	8
Geometry Subtotal		30	25	17
4. Data	4.1. Probability	14	N/A	12
	4.2.Statistics	N/A	12	8
Data Subtotal		14	12	20

The mathematical items were constructed according to a blueprint to represent four standards areas with various competences (concepts and procedures, problem solving, reasoning and communication) and contents (numbers and computation, algebra, geometry, and data with probability and statistics). The blueprint has a hierarchical structure of benchmarks and indicators within standards as shown on Appendix A.

6.2.2 Study 1 Procedure

IRTPRO Beta 2 program was used for the analyses in item difficulty, FICFA, and bifactor analysis. IRTPRO is the most recent IRT application program for item calibration and ability estimation which can provide better estimates for test structure than the other IRT programs (e.g., BILOG, TESTFACT).

6.2.2.1 Study 1-1 Procedure: Item difficulty

For item difficulty, both *b*-parameter value of 2PL IRT model and *p*-value in CTT (the percent of correct answers) were obtained for each grade (see Table 6.2).

Table 6.2 *Item Difficulties of the Four Standards and the Six Benchmarks*

Grade	*Number of items*	*IRT*	*CTT*
6th grade	86 items	*b*-parameter of 2PL model	*p*-value $\left(\frac{\#\ of\ correct\ answers}{\#\ of\ total\ responses}\right)$
7th grade	84 items		
8th grade	86 items		

Then, the average *b*-value and *p*-value of the items involving each standard or each benchmark were considered as the content difficulty. The difficulty of each standard

(benchmark) was compared to see which standard (benchmark) is harder or easier within grades. Note that the higher *b*-value indicates the harder item while the higher *p*-value designates the easier item.

As noted above, it was hypothesized that the difficulties of the four standards would be in the following order: Standard 3 (Geometry) > Standard 2 (Algebra) > Standard 4 (Data) > Standard 1 (Number and computation). Likewise, the difficulties of the six specific benchmarks were compared within grades. It was hypothesized that difficulties of the six specific benchmarks would be in the following order: Benchmark 3.4 (geometry from an algebraic perspective) > Benchmark 3.1 (geometric figures and their properties) > Benchmark 2.2 (variables/equations and inequalities) >Benchmark 4.1 (probability) > Benchmark 1.4 (computation) > Benchmark 1.1 (number sense).

6.2.2.2 Study 1-2 Procedure: Full Information Confirmatory Factor Analysis

First, the four-factor structure defined by the four standards was examined. Second, FICFA model without split loadings (non-zero loadings on only one factor) were compared with the FICFA model with split loadings (non-zero loadings on more than one factor). In this study, the FICFA model without split loadings is based on the blueprint while the FICFA model allowing split loadings is based on the Q matrix constructed by mathematical experts for multiple involvements of standards in items. As a result, three models were tested: Model 1 (single factor model), Model 2 (FICFA model without split loadings), and Model 3(FICFA model with split loadings). Therefore, a total of nine models (with and without split loadings for the three grades) were tested as shown in the Table 6.3.

Table 6.3 *Single Factor Model, FICFA Model without split loading, and FICFA Model with split loadings*

Model	***Factors***	***Data***
#1-1	Single factor model	6^{th} grade
#1-2	Single factor model	7^{th} grade
#1-3	Single factor model	8^{th} grade
#2-1	FICFA model without split loadings (4 Standards)	6^{th} grade
#2-2	FICFA model without split loadings (4 Standards)	7^{th} grade
#2-3	FICFA model without split loadings (4 Standards)	8^{th} grade
#3-1	FICFA model with split loadings (4 Standards)	6^{th} grade
#3-2	FICFA model with split loadings (4 Standards)	7^{th} grade
#3-3	FICFA model with split loadings (4 Standards)	8^{th} grade

Because Model 1, 2, and 3 are nested models, hierarchical chi-square tests were used. To examine the four factor structure of the mathematical test, significance of the chi-square change using the likelihood-based value (-2loglikelihood) between Model 1 and MODEL 2 was tested as follows:

$$\Delta\chi^2 = \text{-2log Likelihood}_{\text{ single factor}} - (\text{ -2log Likelihood}_{\text{ FICFA without split loadings}})$$

To compare MODEL 2 and MODEL 3, the significance test of the chi-square change was as follows:

$$\Delta\chi^2 = \text{-2log Likelihood}_{\text{ without split loadings}} - (\text{ -2log Likelihood}_{\text{ with split loadings}})$$

Factor loadings and goodness of fit statistics such as Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were also compared among MODEL 1, 2, and 3.

6.2.2.3 Study 1-3 Procedure: Bifactor Analysis by Contents

A total of six bifactor models were tested as shown in Table 6.4. It should be noted that Benchmarks 3.4 and 4.1 were not measured in 7th grade, thus four benchmarks (1.1, 1.4, 2.2, and 3.1) were included in the grade. As shown in the table, it was tested whether a general factor (i.e., mathematical competency) and content-specific group factors (i.e., four standards or six benchmarks) exist across the grades. Item bifactor analyses by the four standards and by the six specific benchmarks (1.1, 1.4, 2.2, 3.1, 3.4 and 4.1) were conducted for each grade.

Table 6.4 *Bifactor Models by Standards and Benchmarks within Grades*

Bifactor model	***Number of Group Factors***	***Data***
#1	4 Standards	6th grade
#2	4 Standards	7th grade
#3	4 Standards	8th grade
#4	6 Benchmarks (1.1, 1.4, 2.2, 3.1, 3.4, & 4.1)	6th grade
#5	4 Benchmarks (1.1, 1.4, 2.2, & 3.1)	7th grade
#6	6 Benchmarks (1.1, 1.4, 2.2, 3.1, 3.4, & 4.1)	8th grade

To test the bifactor model, significance of the chi-square between unidimensional model and bifactor model was tested using the likelihood-based value (-2loglikelihood) as follows:

$$\Delta\chi^2 = \text{-2log Likelihood}_{\text{unidimensional}} - (\text{-2log Likelihood}_{\text{bifactor}})$$

Factor loadings as well as goodness of fit statistics such as Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were also examined. In a bifactor model, the evidence of the existence of a general factor is that all items should show high loadings on it. Therefore, the factor loadings can be also used for checking the model fit regarding the general factor. An acceptable fit suggests that a general factor can account for responses on all test items and that the group factors are independent of the general factors (Chen, West, & Sousa, 2006). Additionally, the bifactor models have constraints of no split factor loadings on the content-specific group factors (e.g., four standards) and of no inter-factor correlations. Such constraints may decrease the model-data fit because some test items involve multiple contents (standards, benchmarks) and the group factors may correlate with each other.

6.3 Study1 Results

6.3.1 Study 1-1: Item Difficulty

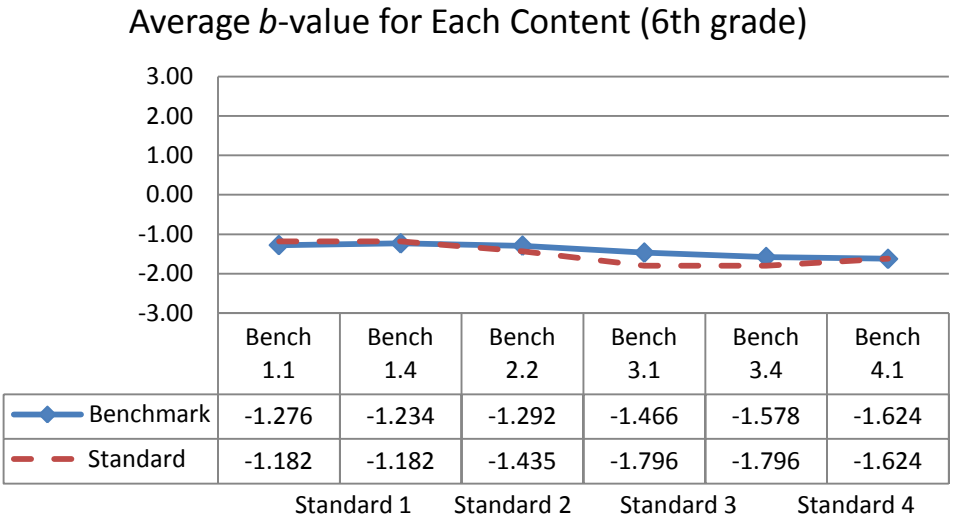
First, item difficulties (*b*- and p-values) of all the test items (86 items for 6^{th} grade, 84 items for 7^{th} grade, and 86 items for 8^{th} grade) were estimated for each grade. Based on item-model fit (summed -score based χ^2), 26 items for 6^{th} grade, 24 items for 7^{th} grade, and 26 items for 8^{th} grade were excluded ($p < .05$), thereby resulting in 60 items for each grade. The results after excluding using item-model fit are shown in Table 6.5. As the content (i.e., standard or benchmark) difficulty, the average *b*-and *p* -values of the items involving each of the four standards and the six benchmarks (1.1, 1.4, 2.2, 3.1, 3.4, and 4.1) were presented for each grade in Figure 6.2, 6.3, and 6.4 respectively. It should be noted that the first four benchmarks (1.1, 1.4, 2.2, and 3.1) were commonly measured in

all the three grades, but the last two benchmarks (3.4 and 4.1) were measured only in 6th and 8th grade.

Table 6.5 *Results after excluding using Item-Model Fit*

Standard	***Benchmark***	***6th Grade***	***7th Grade***	***8th Grade***
Total Objective Questions		86-26=60	84-24=60	86-26=60
1. Number and Computation	1.1. Number sense	10-5=5	5-1=4	6-1=5
	1.2.Number systems and their properties	N/A	N/A	10-1=9
	1.3. Estimation	6-3=3	N/A	N/A
	1.4. Computation	14-4=10	13-6=7	12-5=7
Number and Computation Subtotal		30-12=18	18-7=11	28-7=21
2.Algebra	2.1. Patterns	4-2=2	10-2=8	N/A
	2.2. Variables, Equations, and Inequalities	8-2=6	19-5=14	12-7=5
	2.3. Functions	N/A	N/A	5-3=2
	2.4. Models	N/A	N/A	4
Algebra Subtotal		12-4=8	29-6=22	21-10=11
3.Geometry	3.1. Geometric Figures and Their Properties	6-1=5	7-1=6	9-5=4
	3.2. Measurement and Estimation	14-5=9	14-5=9	N/A
	3.3. Transformational Geometry	4	4-3=1	N/A
	3.4. Geometry from an Algebraic Perspective	6	N/A	8-1=7
Geometry Subtotal		30-6=24	25—9=16	17-6=11
4. Data	4.1. Probability	14-4=10	N/A	12-3=9
	4.2.Statistics	N/A	12-1=11	8
Data Subtotal		14-4=10	11	20-3=17

The results in *b*-value and *p*-value were similar as shown in the three figures (Figure 6.2 for 6th grade, Figure 6.3 for 7th grade Figure 6.4 for 8th grade); there was no significant content difficulty in any grade level.



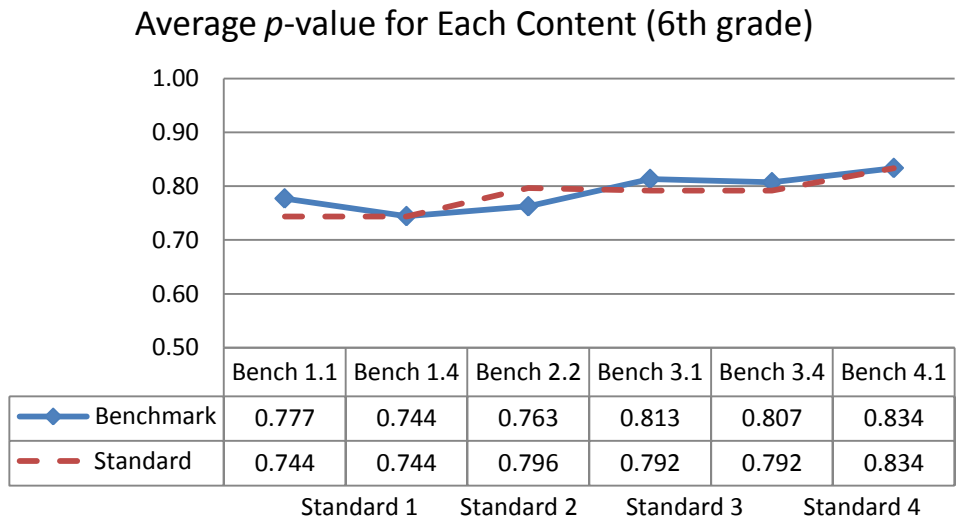


Figure 6.2 Content Difficulties in 6th Grade

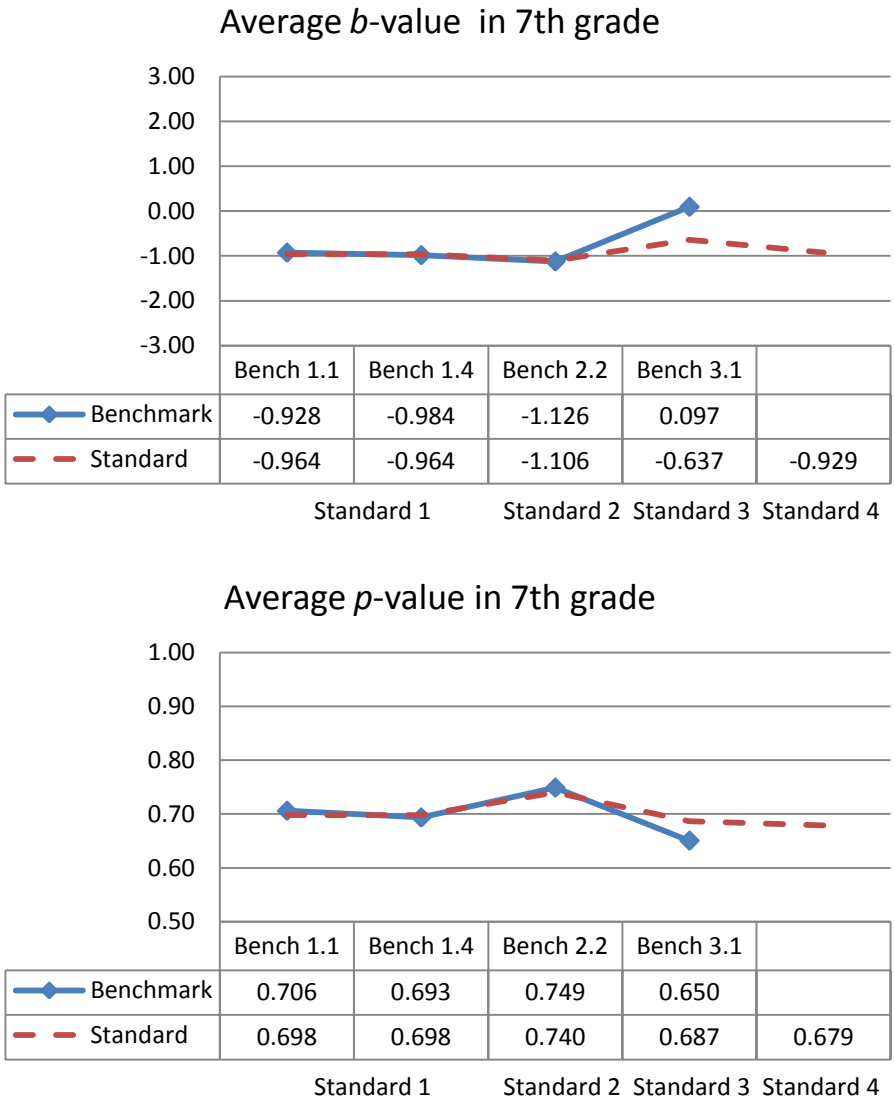


Figure 6.3 Content Difficulties in 7th Grade

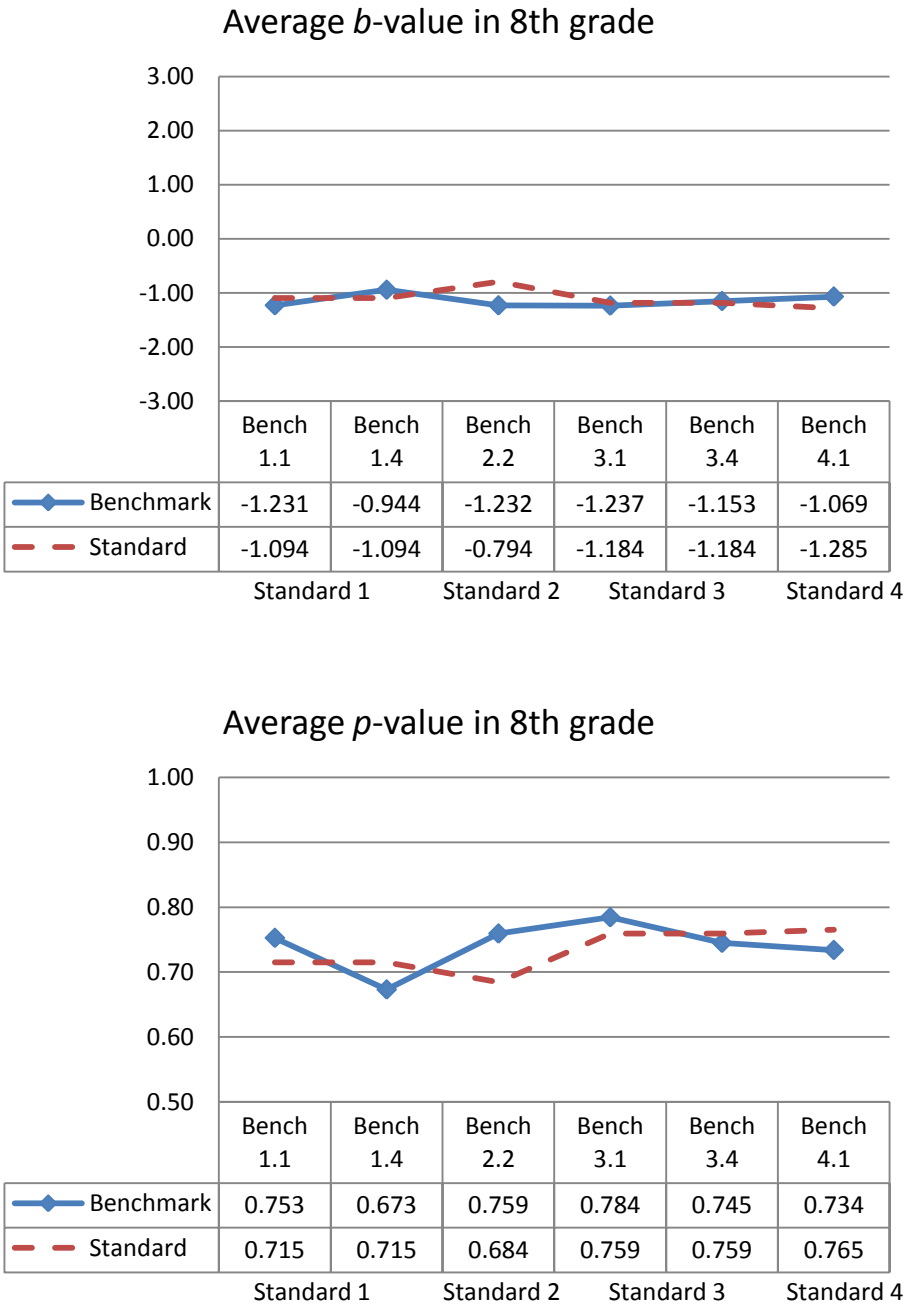


Figure 6.4 Content Difficulties in 8th Grade

Unlike the hypotheses described above regarding the content difficulty, no significant difference was found for the four standards [F (3,56) = 1.34, p = .27 in 6th grade, F (3,56) = 0.51, p = .68 for 7th grade, and F (3,56) = 1.01, p = .40 for 8th grade] and the benchmarks [F (5,36) = 0.63, p = .68 for 6th grade, F (3,27) = 1.15, p = .35 for 7th grade, and F (5,32) = 0.14, p = .98 for 8th grade] in the average b-value. Although the difference was not significant, it was an interesting findings that, Standard 1 (Number and Computation) and Benchmarks 1.1 and 1.4 were generally harder than the other standards or benchmarks in 6th grade while Standard 3 (Geometry) and Benchmark 3.1 (geometric figures and their properties) were hardest in 7th grade. In 8th grade, Standard 2 (Algebra) was the hardest of the four standards. A possible reason that these standards were harder than the others in specific grades is that they were especially targeted in those grades; Standard 1 (Number and computation) may be targeted in 6th grade because children's development of central conceptual structure for number is until ten years old (Case, 1985). Standard 3 (Geometry) may not be emphasized in 6th grade but rather in 7th grade or later because it requires logical reasoning and other mathematical thinking skills (Tatsuoka et al., 2004).

6.3.2 Study 1-2: Full Information Confirmatory Factor Analysis

First, chi-square statistics of the single factor model (Model 1) and the FICFA model (by the four standards) without split loadings (Model 2) were presented for each grade in Table 6.6 along with the significance of the change. All the chi-square changes in the three grades ($\Delta\chi^2 = \chi^2_{\text{single factor}} - \chi^2_{\text{FICFA}}$) were significant, which indicated the four factor model greatly improved the data fit.

Second, as shown in the Table 6.7, two models were tested in each grade: FICFA model without split loadings (Model 2) and FICFA model with split loadings (Model 3). For 6th grade, the chi-square change ($\Delta\chi^2 = \chi^2_{\text{FICFA without split loadings}} - \chi^2_{\text{FICFA with split loadings}}$) was significant, which indicated that Model 3 did significantly improve the data fit. Also, for 7th and 8th grade, it was very significant, which indicated the split loadings of the FICFA model greatly improved the data fit. Also, AIC and BIC also supported that Model 3 had a better fit in all 6th, 7th, and 8th grades. The fit indices improved in the all three grades as shown in Table 6.7.

Table 6.6 *Comparison of the Single Factor Model (Model 1) and the FICFA model without Split Loadings (Model 2) by the Four Standards in Fit Indices*

Grade	*Model*	*-2logL*	$\Delta\chi^2$	Δdf	*P*	*AIC*	*BIC*
6th	#1-1	149285.04				149405.04	149758.36
	#2-1	137875.05	11409.99	60	.000	138115.05	138821.70
7th	#1-2	172330.55				172450.55	172803.88
	#2-2	164698.62	7631.93	60	.000	164938.62	165645.26
8th	#1-3	165310.04				165430.04	165783.37
	#2-3	157195.47	8204.57	60	.000	163514,94	164345.25

Note. #1 = Single factor model, #2 = FICFA model without split loading

Table 6.7 *Comparison of the FICFA model without Split Loadings (Model 2) and the FICFA model with Split Loadings (Model3) in Fit Indices*

Grade	Model	-2logL	$\Delta\chi^2$	Δdf	P	AIC	BIC
6th	#2-1	137875.05				138115.05	138821.70
	#3-1	137798.57	76.48	3	<.0001	138044.57	138768.88
7th	#2-2	164698.62				164938.62	165645.26
	#3-2	163232.94	1465.68	21	.000	163514.94	164345.25
8th	#2-3	157105.47				163514.94	164345.25
	#3-3	154461.00	2644.47	29	.000	154759.00	155636.41

Note. #2 = FICFA model without split loading, #3 = FICFA model with split loading

Factor loadings of all the FICFA models with split loadings for three grades were provided in Appendix B. In 6th grade, all the factor loadings were strong except item # 63 (= 0.18) for the primary factor (Standard 1), item #65, 66, and 67 (0.19, 0.21, and 0.21, respectively) for the secondary factor (Standard 4). In 7th and 8th grades, some items showed low loadings (< 0.25) on the primary factors but high loadings (> 0.25) on the secondary factors (items # 34 for 7th grade, items #24, 51, and 95 for 8th grade). Item #34 in 7th grade involved Standard 2 (Algebra) as a primary factor and Standard 1 (Number and Computation) as a secondary factor. Item #24 in 8th grade involved Standard 4 (Data) as a primary factor and Standard 2 as a secondary factor. Item #51 and #95 in 8th grade were loaded on Standard 3 (Geometry) as a primary factor and Standard 2 as a secondary factor. It was speculated that these reversed factor loading patterns between primary and secondary factor may be due to the more difficult aspect of the secondary factor than the primary factor. For example, in item #55 in 8th grade, the secondary factor (Algebra) was harder than the primary factor (Data) and it may caused the reversed factor loading pattern. Also, item #68 in 7th grade and item #88 in 8th grade showed very low loadings on their primary factors (0.08 on Standard 3 and 0.17 on Standard 1, respectively), which suggested that these items did not represent the factors very well. However, rest of the factor loadings were strong and fit the Q-matrix well.

6.3.3 Study 1-3: Bifactor Analysis by Content

First, chi-square statistics of the unidimensional and bifactor model and the significance of the change between the two models were presented for the standards and the benchmarks in Table 6.8 and 6.9, respectively.

Table 6.8 *Chi-Square Change in the Bifactor Models by Four Standards within Grades*

Grade	*Model*	*-2logL*	$\Delta\chi^2$	Δdf	*p*	*AIC*	*BIC*
6th	Uni	131677.84				131677.84	132624.84
	#1	129948.37	1729.47	60	<.0001	130294.37	131313.12
7th	Uni	156822.59				157062.59	157769.23
	#2	154774.30	2048.29	60	<.0001	155134.30	156194.26
8th	Uni	148967.32				149207.32	149913.97
	#3	147940.41	1026.91	60	<.0001	148300.41	149360.38

Note. Uni = Unidimesional model, #1 = Bifactor model #1

Table 6.9 *Chi-Square Change in the Bifactor Models by Six Benchmarks within Grades*

Grade	*Model*	*-2logL*	$\Delta\chi^2$	Δdf	*P*	*AIC*	*BIC*
6th	Uni	93699.82				93867.82	94362.47
	#4	91786.99	1912.83	42	<.0001	92038.99	92780.97
7th	Uni	81981.93				82105.93	82471.03
	#5	81421.20	560.73	31	<.0001	81607.20	82154.85
8th	Uni	95471.33				95623.33	96070.87
	#6	93055.00	2416.33	38	<.0001	93283.00	93954.31

Note. Uni = Unidimesional model, #4 = Bifactor model #4

All the chi-square changes in the three grades ($\Delta\chi^2$ = -2log Likelihood $_{\text{unidimensional}}$ − (-2log Likelihood $_{\text{bifactor}}$)) were very significant, which supports that the general factor (i.e., mathematical competency) and the content-specific group factors (i.e., standards or benchmarks) exist in the tests of all the grades.

Second, factor loadings of the six bifactor models were presented in Appendix C. All the items showed high factor loadings on the general factor in all grades, which strongly suggested the existence of a general factor in the mathematical achievement test in all grades. However, many items showed nearly zero loadings on specific group

factors such as items 15 through 19 and 67 on Standard 1, and items 41 through 50 and 71 though 74 on Standard 3 in 6th grade, which implied that the group factor was not independent of the general factor in such items.

6.4 Study 1 Summary and Discussion

First, there were two hypotheses regarding item difficulty as follows:

Hypothesis 1. Difficulties of the four standards are in the following order: Standard 3 (Geometry) > Standard 2 (Algebra) > Standard 4 (Data) > Standard 1 (Number and computation). * harder >easier

Hypothesis 2. Difficulties of the six specific benchmarks are in the following order: Benchmark 3.4 (geometry from an algebraic perspective) > Benchmark 3.1 (geometric figures and their properties) > Benchmark 2.2 (variables/equations and inequalities) > Benchmark 4.1 (probability) > Benchmark 1.4 (computation) > Benchmark 1.1 (number sense). * harder >easier

Therefore *Hypotheses* #1 and #2 were not supported. However, Benchmark 3.1 (geometric figures and their properties) seemed much more difficult than other benchmarks in 7th grade although it was not statistically significant. As shown in Table 5.10 (Indicators of the Benchmark 3.1 of Standard 3), Benchmark 3.1 in 7th grade requires more complex concepts and knowledge compared with in 6th and 8th grade. That is, it requires how to classify only triangles in 6th grade, but how to identify angles and side properties of triangles and quadrilaterals in 7th grade. In 8th grade, it requires students to use simply Pythagorean theorem to solve application problems. It was speculated that the requirements of Benchmark 3.1 in 7th grade may be very challenging. Other reason is

that the lack of impact on item difficulty could be due to selecting only easy items of this type for the test or heavy emphasis of the area in teaching.

Second, there were two hypotheses regarding FICFA and its split loadings:

Hypothesis 3. The four factor structure will be identified for each grade.

Hypothesis 4. The FICFA model with split loadings fits the data better than the FICFA model without split loadings.

Chi-square tests and fit indices supported the existence of the four-factor structure by the four standards in the mathematical achievement test in all the three grades. Hypothesis 4 was well supported because the split loadings (involvement of items in multiple standards) were very significant in all grades. One interesting finding in the FICFA models with split loadings was that some items showed reversed factor patterns such as high loadings on the secondary factors and low loadings on the primary factors. The finding may suggest that these reversed factor loading patterns between primary and secondary factor may be due to the more difficult aspect of the secondary factor than the primary factor.

Third, following was the hypothesis regarding bifactor model:

Hypothesis 5. The bifactor structure which includes a general factor (mathematical competency) and content-specific group factors (four standards or six benchmarks) exists across the grades.

Significant chi-square changes and high factor loadings of all the items on the general factor strongly supported the existence of a general factor in the mathematical achievement test in all grades. Therefore, Hypothesis 5 was supported. Additionally, many items showed nearly zero loadings on some group factors (standard or benchmark),

which suggested that group factors were not independent of the general factor. In other words, the finding supported that the general factor was sufficient to account for the correlations between such items as having nearly zero group factors loadings.

CHAPTER 7

STUDY 2: MEASURING MATHEMATICAL ACHIEVEMENT CHANGE

7.1 Study 2 Purpose and Hypotheses

7.1.1 Study 2 Purpose

The main purpose of this study was to examine how middle school students change on mathematical achievement from 6th grade to 8th grade. Four standards and the six specific benchmarks (1.1, 1.4, 2.2, 3.1, 3.4, and 4.1) that were examined in Study 1 were used for assessing the change. Based on Piaget's cognitive development stages, 6th graders can be in either concrete operational stage (from age 6 or 7 to age 11 or 12) or formal operational stage (age 11 or 12 and up) while 7th and 8th graders should be in formal operational stage because the math achievement test in this study was administered at the end of every school year. In this longitudinal study, the middle school students' mathematical achievement changes were measured and analyzed using the perspective of the two cognitive development theories (Piaget's developmental stages and Vygotsky's social-cultural theory). Additionally, it was examined if gender and socioeconomic status (SES) affect the mathematical achievement in each grade and the changes over the three grades. A consistent association has been found between SES and mathematical Achievement over the years. SES may have many aspects of definition. National Mathematics Advisory Panel (NMAP; 2008) included parental education, poverty level, parental income, or a composite index to define SES. Therefore, SES differences may be used as a supporting evidence of the Vygotsky's theory that children can improve their school achievement when mentoring is available.

7.1.2 Study 2 Hypotheses

First, Muzzatti and Agnoli (2007) claimed that the gender difference in children's mathematical performance was not biologically determined but socioculturally determined. They claimed that there was a gender difference in children's attitudes toward mathematics and the stereotype that boys were better than girls in mathematics limited the career chances in mathematics for females in the future. Actually, the gender difference in mathematics has long been an issue. Much research has been done on the issue and past studies have shown that males outperformed females in mathematical achievement (Hopkins, 2004). However, a majority of the current studies conclude that the gender gap in mathematics has been decreasing in recent decades and is now quite small (Fennema, 1996; Gray, 1996; Hanna, 2003; Leahy & Guo, 2001; Wellesley College, 1992). Interestingly, girls even outperform boys in mathematics achievement at the elementary and middle school ages (Ansell & Doerr, 2000; Fennema, 1976; Friedman, 1989; Sprigler & Alsup, 2003), but, at the high school ages, differences in mathematical performance tend to favor males in the widely used college entrance exams, the ACT and SAT (Hopkins, 2004). Ansell and Doerr (2000) reported that no statistically significant gender difference existed for overall average scores in the NAEP (National Assessment of Educational Progress) data, but males significantly outperformed females in measurement, geometry and spatial sense. Also, National Mathematics Advisory Panel (NMAP; 2008) reported that the average mathematical achievement of boys and girls showed very similar levels (in favor of boys) in large nationally representative samples.

Second, it is also a well known issue that low SES is linked to poor mathematical performance (NMAP, 2008). Based on the NAEP data in 2001 and 2002, the

achievement gaps between the high SES and low SES students existed regardless of race (Hopkins, 2004). Guo (1998) claimed that poverty has a significant negative effect on the cognitive development in childhood while it has an effect on achievement measures in adolescence. In some studies, it was suggested that a more important factor than SES is the home environment factor such as the number of stimulating toys, quality of child-parent relationship (Crane, 1996), inadequate social experiences and learning opportunities in academic achievement (NMAP, 2008).

Third, most 6^{th} graders are 11 to 12 years old, thus 6^{th} grade students can be in either concrete operational (age 6 or 7 until age 11 or 12) stage or formal operational stage (age 11 or 12 through adulthood). In concrete operational stage, although they have advancements in reasoning, they seemed to deal with concrete observable objects only, thus still having limitation for applying their logical operations (Ormrod, 2008). However, when they enter formal operations stage, they are able to think and reason about things that have little or no basis in physical reality such as abstract concepts, hypothetical ideas, and contrary-to-fact statements (Ormrod, 2008). It can be assumed that 7^{th} and 8^{th} graders are in formal operations stage (age 11 or 12 through adulthood) because 7^{th} graders are normally 12 to 13 years old and 8^{th} graders are normally 13 to 14 years old. Therefore, it is expected that a significant change can be observed in the mathematical achievement between 6^{th} grade and the next two grade levels (7^{th} and 8^{th}).

As a result, following hypotheses were made regarding gender, SES, and age in mathematical achievement. Two contradictory hypotheses regarding the gender difference will be tested as follows:

Hypothesis 6-a. There is no significant gender difference in mathematical achievement between female and male.

Hypothesis 6-b. Males outperform females in the mathematical achievement test.

*Hypothesis 7.*There is a significant SES difference in mathematical achievement.

Hypothesis 8. There is a significant incremental difference in mathematical achievement between 6^{th} grade and the next two grade levels (7^{th} and 8^{th}); a quadratic trend of growth exists, such that the change from 6^{th} to 7^{th} grade is greater than from 7^{th} to 8^{th} grade.

To test these hypotheses, three models were used: full information CFA (FICFA) model, bifactor model, and MRMLC (Embretson, 1991). Although FICFA and bifactor model were not originally designed for a longitudinal study, it was explored how to apply these models to the longitudinal data in addition to MRMLC application. FICFA and bifactor analysis were conducted using three time factors (6^{th}, 7^{th}, and 8^{th} grades) instead of the four standards which were used as four factors in Study 1. In the FICFA model, the examinee competency levels (θ's) on the three occasions (6^{th}, 7^{th}, and 8^{th} grades) were estimated. In a bifactor model, factor loadings on specific group factors may indicate the extent to which the group factors have effect on the general factor. Therefore, the factor loadings on the three time factors were estimated and compared to see which occasion had bigger effect on the general factor. Second, MRMLC is one of the available psychometric models for longitudinal studies of ability or measuring learning potential assessment. One important advantage of using MRMLC in this study is that this model was developed for situations in which items are not repeated as the data in this study, thereby avoiding practice (or memory) effects and local dependency among item

responses when items are repeated (von Davier & Xu, 2009). The simultaneous estimation of MRMLC parameters is available only recently as a generalized procedures for confirmatory multidimensional IRT models in IRTPRO was developed.

7.2 Study 2 Method

7.2.1 Subjects and Instruments

Same mathematical achievement test data of same subjects as in Study 1 was used for this study. The four mathematical standards and the six specific benchmarks were used for assessing the achievement changes from 6^{th} to 8^{th} grade. The six specific benchmarks include Benchmarks 1.1 (number sense), 1.4 (computation), 2.2 (variables/equations and inequalities), 3.1 (geometric figures and their properties), 3.4 (geometry from an algebraic perspective), and 4.1 (probability). It should be noted that only items involving each standard and each benchmark was used for the analysis.

Of the total 2667 students, 1360 (51%) were males and 1307 (49%) were females. For race, 1816 (68.1%) were Whites, 369 (13.8%) were Hispanics, 271 (10.2%) were African Americans, and 211 (7.9%) were others (Asians, Native American, multi and missing). With regard to the national school lunch program, 1580 (59.2%) paid regular prices for lunch and 1087 (41.7%) students were provided free or reduced price lunch. Of the regular price lunch group, 1325 (84%) were Whites, 79 (5%) were Hispanics, 66 (4%) were African Americans, and 110(7%) were others (Asians, American Indians, and multi or missing). Of the free and reduced price lunch group, 491 (45%) were Whites, 290 (27%) were Hispanics, 205 (19%) were African Americans, and 101 (9%) were others. Figure 7.1 presents these proportions of races in each of the two lunch groups. Also, cross

tabulations of gender × school lunch program and race × school lunch program are provided in Appendix D. For the SES group classification, school lunch program information was used. Students were classified into two SES groups: (1) regular price lunch group (= Reg) and (2) free/reduced price lunch group (= F/R).

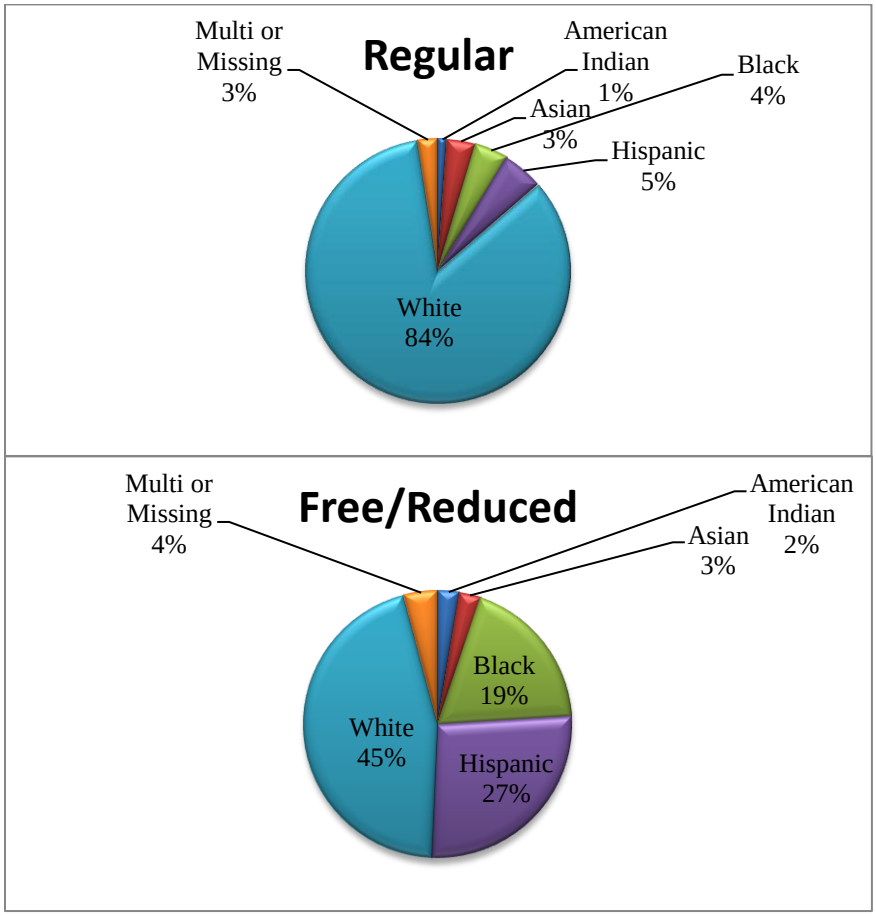


Figure 7.1 Proportions of Races in the Regular Price Lunch Group and Free/Reduced Price Lunch group.

7.2.2 Study 2 Procedure

All the analyses in this study (FICFA, bifactor, and MRMLC analyses) were conducted using IRTPRO Beta 2 program. Although FICFA model and bifactor model

were not originally designed for measuring the ability changes over time, it was explored how to apply the models to the longitudinal data in addition to MRMLC application.

7.2.2.1 Study 2-1 Procedure: FICFA and Bifactor Analysis and by Time Factors

FICFA and bifactor models by the three time factors (6^{th}, 7^{th}, and 8^{th} grades) were tested for each of the ten contents (4 standards and 6 benchmarks) as shown in Table 7.1. Note that each model includes only items involving the corresponding content, thus having a different number of items in it. Same content and data were used for the ten FICFA models (Model #1 through #10) and the ten bifactor models (Model #1through #10).

Table 7.1 *FICFA and Bifactor Models by Time Factors (6^{th}, 7^{th}, & 8^{th} grade)*

Model	***Content***	***Data***
#1	Standards 1	50 Items involving Standard 1 of all grades
#2	Standards 2	41 Items involving Standard 2 of all grades
#3	Standards 3	51 Items involving Standard 3 of all grades
#4	Standards 4	38 Items involving Standard 4 of all grades
#5	Benchmark 1.1	14 Items involving Benchmark 1.1 of all grades
#6	Benchmark 1.4	24 Items involving Benchmark 1.4 of all grades
#7	Benchmark 2.2	25 Items involving Benchmark 2.2 of all grades
#8	Benchmark 3.1	15 Items involving Benchmark 3.1 of all grades
#9	Benchmark 3.4	13 Items involving Benchmark 3.4 of 6^{th} and 8^{th}
#10	Benchmark 4.1	20 Items involving Benchmark 4.1 of 6^{th} and 8^{th}

First, in the FICFA model (see Figure 7.2-A), each factor was regarded as the competency of the content at each of the three occasions (6^{th}, 7^{th}, and 8^{th} grade). Examinees' competency levels (θ's) on each occasion was estimated using 2-PLM. It should be noted that each estimated θ is not the absolute competency level of the

examinee but the relative competency level in its grade only since it is the repeated measured longitudinal data. Therefore, the result provided the information regarding the relative mathematical achievement change across the three grades. On each occasion of each content, different gender and SES groups were compared in their competency levels.

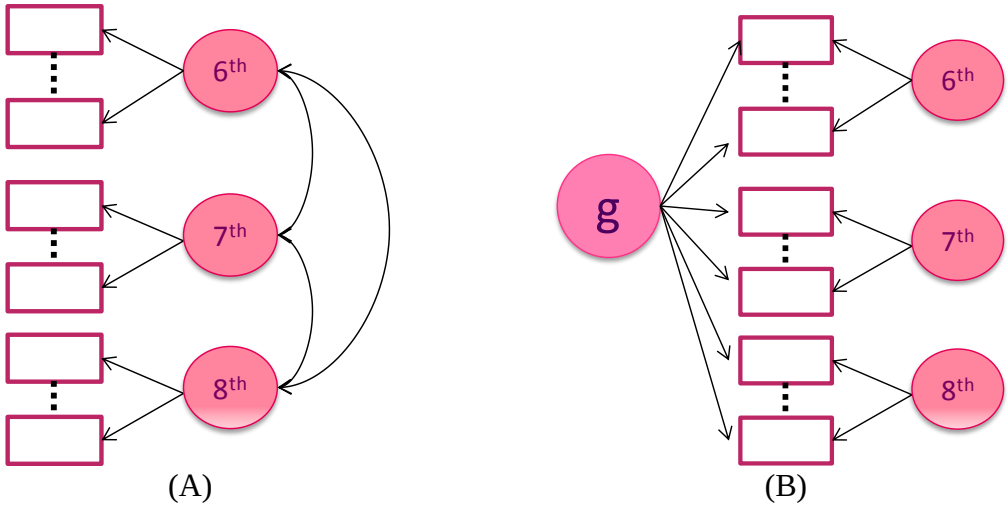


Figure 7.2 FICFA Model (A) and Bifactor Model (B) by Three Time Factors

Second, in the bifactor model, the general factor can be counted as the overall competency of the content while the group factor may represent the unique competency of the content at the specific time (6^{th}, 7^{th}, or 8^{th} grade). The bifactor model was graphically presented in Figure 7.2-B. It was assumed that the three time factors were uncorrelated with each other according to the constrain of the bifactor model. Factor loadings on the time-specific factors were examined to see which occasion had bigger effect on the general factor.

7.2.2.2 Study 2-2 Procedure: MRMLC Analysis

MRMLC (Embretson, 1991) is a longitudinal psychometric model which is appropriate for measuring learning and change over occasions. Therefore, in this study, MRMLC was used to measure the mathematical achievement change from 6^{th} to 8^{th} grade. (see Table 7.2). The four standards and the four specific benchmarks were used as a dependent variable of each MRMLC model in Table 7.2.

Table 7.2 *Measuring Mathematical Achievement Change from 6^{th} to 8^{th} grade*

MRMLC	***Content***	***Data***
#1	Standards 1	50 Items involving Standard 1 of all grades
#2	Standards 2	41 Items involving Standard 2 of all grades
#3	Standards 3	51 Items involving Standard 3 of all grades
#4	Standards 4	38 Items involving Standard 4 of all grades
#5	Benchmark 1.1	14 Items involving Benchmark 1.1 of all grades
#6	Benchmark 1.4	24 Items involving Benchmark 1.4 of all grades
#7	Benchmark 2.2	25 Items involving Benchmark 2.2 of all grades
#8	Benchmark 3.1	15 Items involving Benchmark 3.1 of all grades
#9	Benchmark 3.4	13 Items involving Benchmark 3.4 of 6^{th} and 8^{th}
#10	Benchmark 4.1	20 Items involving Benchmark 4.1 of 6^{th} and 8^{th}

MRMLC uses a Wiener process structure which includes the initial trait level and one or more modifiability. In the Wiener process structure, each new occasion involves a new dimension and MRMLC treats modifiability as separate dimensions. For the measurements over the three occasions (6^{th}, 7^{th}, and 8^{th} grade) in this study, the Wiener process structure can be designed as in Table 7.3.

In this table, the columns represent dimensions of trait levels and the rows represent the occasions under which items are observed; 60items for 6^{th} grade, 60 items for 7^{th} grade, and 60 items for 8^{th} grade. The value '1' denotes the condition in which a

particular dimension is involved in performance, while '0' denotes the condition in which a particular dimension is not involved in performance. The initial trait level (θ_1) is involved in all occasions. The second trait level (θ_2) is a modifiability that represents the change from 6th grade to 7th grade. The third trait level (θ_3) is a modifiability that represents the change from 7th grade to 8th grade. Thus, the person ability or achievement at 6th, 7th, and 8th grade will be represented by θ_1, $\theta_1 + \theta_2$, and $\theta_1 + \theta_2 + \theta_3$, respectively. It should be noted that since the item difficulty, b_i, in MRMLC (see Equation 4.13) is assumed to remain constant across occasions, a positive modifiability indicates performance increase (achievement improvement).

Table 7.3 *Wiener Process Structure for the Occasions of 6th, 7th, and 8th Grades*

Occasion	*Dimensions*		
	θ_1 (Initial trait level)	θ_2 (Modifiability)	θ_3 (Modifiability)
6th grade (60 items)	1	0	0
7th grade (60 items)	1	1	0
8th grade (60 items)	1	1	1

Then, a repeated-measures ANOVA was conducted to test the significance of the mathematical achievement change. For a significant change, ANOVA contrast was performed to find where a significant change exists among the three grades. The Helmert contrast in which each level (i.e., grade) was compared to the mean of the subsequent levels was used for the contrast test.

7.3 Study 2 Results

7.3.1 Study 2-1: FICFA and Bifactor Analysis by Time Factor

First, the descriptive statistics of the θ estimates on the three occasions (6^{th}, 7^{th}, and 8^{th} grade) in the FICFA model were reported for each content in Table 7.4. It should be noted that the estimated mean θ in this table is a relative competency level in each grade, thus it cannot be used for measuring the true competency level changes over the three grades. As shown in Table 7.4, 6^{th} graders showed the better performance in Standards 1 (Number and Computation) and 4 (Data) than in the other standards. They showed the lowest competency level in Standard 2 (Algebra) in their level. 7^{th} and 8^{th} graders greatly improved the competency in Algebra, but showed the lower achievement in Standards 1 and 4 in their grade levels.

Table 7.4 *Descriptive Statistics of the θ Estimates of the Three Time Occasions*

Content	6^{th} grade		7^{th} grade		8^{th} grade	
	M	*SD*	*M*	*SD*	*M*	*SD*
Standard 1	0.325	0.681	0.179	0.629	0.136	0.708
Standard 2	0.134	0.713	0.301	0.465	0.300	0.464
Standard 3	0.303	0.813	0.284	0.597	0.300	0.556
Standard 4	0.381	0.731	0.201	0.641	0.168	0.702
Benchmark 1.1	0.224	0.705	0.138	0.681	0.125	0.609
Benchmark 1.4	0.374	0.633	0.135	0.592	0.138	0.607
Benchmark 2.2	0.180	0.705	0.055	0.781	0.105	0.666
Benchmark 3.1	0.215	0.762	0.382	0.539	0.292	0.594
Benchmark 3.4	0.194	0.705	.	.	-0.083	0.765
Benchmark 4.1	0.303	0.647	.	.	0.099	0.765

Second, significant differences ($\alpha = .01$) in mathematical achievement were found for gender in some contents and for SES in all the contents as shown in Table 7.5.

For gender, females performed better than males in Standard 4 (Data) in 6th grade, $F(1, 2665) = 51.19$, $p = .000$, but no significant difference was found in the later grades (7th and 8th), $F(1, 2665) = 1.62$, $p = .203$ and $F(1, 2665) = 2.85$, $p = .091$, respectively. In 6th grade, males outperformed females in Benchmarks 1.1 (number sense), $F(1, 2665) = 12.80$, $p = .000$, Benchmark 1.4 (computation), $F(1, 2665) = 14.34$, $p = .000$, Benchmark 2.2 (variables, equations, and inequalities), $F(1, 2665) = 11.19$, $p = .001$, and Benchmark 3.1 (geometric figures and their properties), $F(1, 2665) = 7.36$, $p = .007$. However, in 7th and 8th grade, the differences in these four benchmarks became non-significant except Benchmark 3.1. Interestingly, females outperformed on Benchmark 3.1 in 7th and 8th grade, $F(1, 2665) = 7.81$, $p = .005$ and $F(1, 2665) = 17.29$, $p = .000$, respectively. It was also an interesting finding that females outperformed males in Standard 3 (Geometry) at 7th and 8th grade, $F(1, 2665) = 11.50$, $p = .001$ and $F(1, 2665) = 9.82$, $p = .002$, respectively, although both genders had no difference at 6th grade, $F(1, 2665) = .89$, $p = .345$. Finally, Males outperformed females in Standard 2 (Algebra) at 8th grade, $F(1, 2665) = 10.08$, $p = .002$ although they made no difference in 6th and 7th grade. To sum up, it appeared that males performed better in Numbers and Computation and Algebra while females performed better in Geometry and Data. However, in 7th and 8th grade, the gender differences became less significant. This finding was consistent with the previous finding in some studies that girls outperformed boys at the middle school ages but, boys began excelling at the high school ages and outperformed in the college entrance exams such as ACT and SAT (Hopkins, 2004).

Table 7.5 *Significant Difference in Mathematical Achievement for Gender and SES*

	6^{th} *grade*		7^{th} *grade*		8^{th} *grade*	
Content	*Gender*	*SES*	*gender*	*SES*	*Gender*	*SES*
Standard 1	M=F	Reg>F/R	M=F	Reg>F/R	M=F	Reg>F/R
Standard 2	M=F	Reg>F/R	M=F	Reg>F/R	M>F	Reg>F/R
Standard 3	M=F	Reg>F/R	M<F	Reg>F/R	M<F	Reg>F/R
Standard 4	M<F	Reg>F/R	M=F	Reg>F/R	M=F	Reg>F/R
Benchmark 1.1	M>F	Reg>F/R	M=F	Reg>F/R	M=F	Reg>F/R
Benchmark 1.4	M>F	Reg>F/R	M=F	Reg>F/R	M=F	Reg>F/R
Benchmark 2.2	M>F	Reg>F/R	M=F	Reg>F/R	M=F	Reg>F/R
Benchmark 3.1	M>F	Reg>F/R	M<F	Reg>F/R	M<F	Reg>F/R
Benchmark 3.4	M=F	Reg>F/R	.	.	M=F	Reg>F/R
Benchmark 4.1	M<F	Reg>F/R	.	.	M=F	Reg>F/R

Note. M(male), F(female), Reg(regular price lunch), F/R(free/reduced price lunch), A>B (A is significantly bigger than B at $\alpha = .01$), = (not significantly different)

For SES, students with regular price lunch (Reg) performed significantly better than students with free/reduced price (F/R) in all the contents and grade levels ($p < .001$). It was an interesting finding that students with regular price lunch group students performed always better than students with free lunch or reduced priced lunch in all the contents. The result strongly supported the well-known fact that low SES students have low mathematical achievement.

Third, the average factor loadings of each bifactor model by three time factors (6^{th}, 7^{th}, and 8^{th} grades) were provided for the content in Table 7.6. All the loadings on the general factor were relatively strong in all the models, which indicated that all the test items commonly represented a general factor (i.e., overall competency of the content beyond the occasions). However, the loadings on the specific time factors (three grades levels) showed small to medium sizes, ranging from -.006 to .662 (see Table 7.6).

Table 7.6 *Average Factor Loadings of the Bifactor Model by the Three Time Factors for Each Content*

Content	General Factor	*6th grade*	*7th grade*	*8th grade*
Standard 1	0.529	0.224	0.153	0.157
Standard 2	0.568	0.285	0.189	0.155
Standard 3	0.525	0.014	0.276	0.360
Standard 4	0.554	0.390	0.219	0.049
Benchmark 1.1	0.535	0.312	0.365	0.166
Benchmark 1.4	0.539	0.190	0.196	0.254
Benchmark 2.2	0.576	0.278	0.225	0.194
Benchmark 3.1	0.521	0.442	0.093	0.478
Benchmark 3.4	0.532	0.662	.	-0.006
Benchmark 4.1	0.573	0.447	.	0.096

In Standards 1 (Numbers and Computation), 2 (Algebra), and 4 (Data) and Benchmarks 2.2 (variables, equations, and inequalities), 3.4 (geometry from an algebraic perspective), and 4.1 (probability), the loadings on 6th grade were higher than the loadings on the other grades, which suggests that the competency level on 6th grade had strongest effect on the overall competency level in these contents. The finding may be explained by the proposition, as mentioned earlier, that children develop their central conceptual structure for number until ten years old (Case, 1985).

On the other hand, in Standard 3 (Geometry), the time factor loadings became bigger for the higher grade levels, which may imply that 8th grade was the most important occasion to obtain the specific competency in geometry. The finding supported the previous hypotheses regarding geometry. As mentioned earlier, Tatsuoka et al. (2004) maintained that the success in geometry was highly associated with logical reasoning and other mathematical thinking skills. Ormrod (2008) claimed that children in concrete operations stage have limitation for applying logical operation. Therefore, it can be

inferred that many mathematical skills obtained during the three grades including logical reasoning are needed in order to improve the geometry competency. However, in Benchmarks 3.4 and 4.1, nearly zero factor loadings were observed on 8th grade, which may suggest that the most competency level of these two contents can be obtained in 6th grade. Overall, the different patterns of loadings on the three time factor showed the distinctiveness of the different contents in the contribution of the three grade levels on the overall competency level.

7.3.2 Study 2-2: MRMLC Analysis

Table 7.7 presents the average trait level at 6th grade (θ_1), modifiabilities at 7th (θ_2) and 8th grade (θ_2) and trait levels at 7th and 8th grade for each content. As mentioned above, the trait levels at 7th and 8th grade can be obtained by $\theta_1 + \theta_2$ and $\theta_1 + \theta_2 + \theta_3$, respectively. Repeated-measures ANOVA's indicated that the trait levels in the three grades were significantly different (p = .000) for all the four standards and the six benchmarks. The Helmert contrasts indicated that the difference between 6th grade and the later two grades (7th and 8th grade) and the difference between 7th and 8th grades were all significant (p = .000) in all the contents. However, the differences between 7th and 8th grades were not significant in some contents for some SES groups, especially, reduced price-lunch group and free lunch group. Average trait level changes from 6th to 8th grade for each of the four subgroups (Female, Male, regular price lunch, and free/reduced price lunch) were provided in Appendix E.

Table 7.7 *Average Mathematical Achievement Change from 6th to 8th grade*

Content	*Trait level at 6th grade* (θ_1)	θ_2	θ_3	*Trait level at 7th grade* ($=\theta_1+\theta_2$)	*Trait level at 8th grade* ($=\theta_1+\theta_2+\theta_3$)
Standard 1	0.210	0.887	0.076	1.097	1.173
Standard 2	0.252	0.872	0.128	1.125	1.253
Standard 3	0.143	0.643	0.520	0.786	1.306
Standard 4	0.536	0.266	0.229	0.802	1.031
Benchmark 1.1	0.185	0.609	0.277	0.794	0.918
Benchmark 1.4	0.121	0.209	0.224	0.330	0.554
Benchmark 2.2	0.179	0.996	0.166	1.175	1.340
Benchmark 3.1	0.167	0.623	0.147	0.791	0.938
Benchmark 3.4	0.286	.	0.648	.	0.934
Benchmark 4.1	0.193	.	0.779	.	0.972

7.3.2.1 Standard 1

For overall data in Standard 1 (Number and Computation), the two Helmert contrasts were significant: (1) 6th grade versus later [F (1,2666) = 2282.17, p = .000, η_p^2 = .673] and (2) 7th versus 8th grade [F (1,2666) = 46.24, p = .000, η_p^2 = .017]. Also the two contrasts were also significant (p = .000) for male, female, and regular price-lunch group, However, the second contrast (7th versus 8th) was not significant for the free and reduced price-lunch group, F(1,1086) = 0.923, p = .337, η_p^2 = .001, which indicated that there was no improvement at 8th grade for the low SES group. Thus, no gender difference but a SES difference was found in the achievement change of number and computation. The significance test results of the two contrasts and the achievement changes over the three grades for each of the five groups are presented in Table 7.8 and Figure 7.3.

Table 7.8 *Significance of the Mathematical Achievement Change from 6th to 8th grade in Standard 1*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6th vs. later	*F*(1,2666) = 5478.837	*p* = .000	.673
	7th vs. 8th	*F*(1,2666) = 46.239	*p* = .000	.017
Female	6th vs. later	*F*(1,1359) = 3353.922	*p* = .000	.712
	7th vs. 8th	*F*(1,1359) = 17.046	*p* = .000	.012
Male	6th vs. later	*F*(1,1306) = 2244.953	*p* = .000	.632
	7th vs. 8th	*F*(1,1306) = 22.986	*p* = .000	.022
Regular	6th vs. later	*F*(1,1579) = 3668.296	*p* = .000	.699
	7th vs. 8th	*F*(1,1579) = 87.425	*p* = .000	.052
Free/Reduced	6th vs. later	*F*(1,1086) = 1906.131	*p* = .000	.637
	7th vs. 8th	*F*(1,1086) = 0.923	***p* = .337**	.001

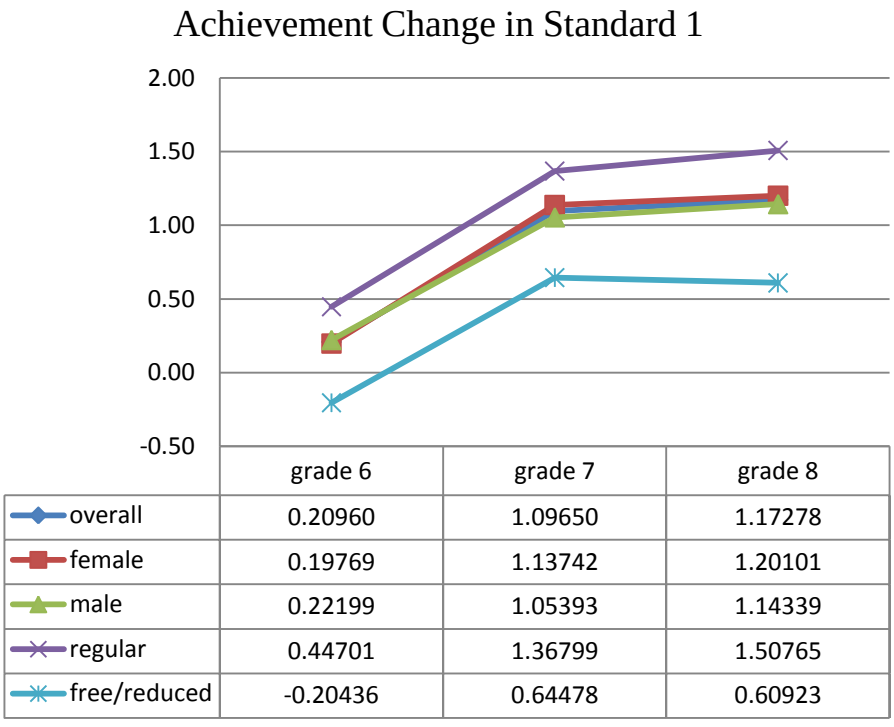


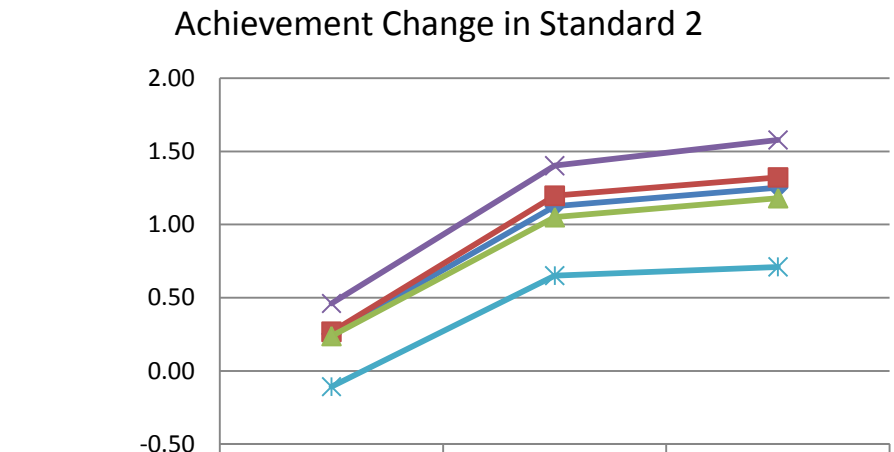
Figure 7.3 Achievement Change from 6th to 8th Grade in Standard 1 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.2 Standard 2

For the achievement change in Standard 2 (Algebra), the two Helmert contrasts were significant in the overall data: (1) 6th grade versus later [F (1,2666) = 3824.30, p = .000, η_p^2 = .589] and (2) 7th versus 8th grade [F (1,2666) = 157.45, p = .000, η_p^2 = .056]. Also, for all the four subgroups (male, female, regular price lunch, free/reduced price lunch), the two contrasts were significant, which indicated that the algebra achievement was significantly improved at 7th and 8th grade in all groups. Therefore, there was no gender and SES difference found in the achievement change in algebra. Table 7.9 and Figure 7.4 presents the significance test results of the two contrasts and the achievement changes for the five groups.

Table 7.9 *Significance of the Mathematical Achievement Change from 6th to 8th grade in Standard 2*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6th vs. later	F(1,2666) = 3824.297	p = .000	.589
	7th vs. 8th	F(1,2666) = 157.449	p = .000	.056
Female	6th vs. later	F(1,1359) = 2265.398	p = .000	.625
	7th vs. 8th	F(1,1359) = 77.921	p = .000	.054
Male	6th vs. later	F(1,1306) = 1609.006	p = .000	.552
	7th vs. 8th	F(1,1306) = 79.505	p = .000	.057
Regular	6th vs. later	F(1,1579) = 2818.636	p = .000	.641
	7th vs. 8th	F(1,1579) = 173.250	p = .000	.099
Free/Reduced	6th vs. later	F(1,1086) = 1150.264	p = .000	.514
	7th vs. 8th	F(1,1086) = 14.097	p = .001	.013



	grade 6	grade 7	grade 8
overall	0.25239	1.12480	1.25288
female	0.26702	1.19662	1.32279
male	0.23717	1.05007	1.18014
regular	0.46067	1.40197	1.57798
free/reduced	-0.10691	0.65148	0.71093

Figure 7.4 Achievement Change from 6th to 8th Grade in Standard 2 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.3 Standard 3

In Standard 3 (Geometry), a clear linear trend in the achievement changes over the three grades was found for overall data, F (1,2666) = 4592.66, p = .000, η_p^2 = .633 as well as for the four subgroups (male, female, regular price lunch, free/reduced price lunch). Also, the two Helmert contrasts (6th grade versus later, 7th versus 8th grade) were significant for overall data and the four subgroups. Therefore, neither gender difference nor SES difference was found in the achievement changes in geometry. The significance test results of the two contrasts and the achievement changes over the three grades for each of the five groups are presented in Table 7.10 and Figure 7.5.

Table 7.10 *Significance of the Mathematical Achievement Change from 6th to 8th grade in Standard 3*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6th vs. later	$F(1,2666) = 4206.930$	$p = .000$	.612
	7th vs. 8th	$F(1,2666) = 2073.201$	$p = .000$	.437
Female	6th vs. later	$F(1,1359) = 2363.109$	$p = .000$	.635
	7th vs. 8th	$F(1,1359) = 1026.182$	$p = .000$	.430
Male	6th vs. later	$F(1,1306) = 1866.941$	$p = .000$	.588
	7th vs. 8th	$F(1,1306) = 1047.407$	$p = .000$	.445
Regular	6th vs. later	$F(1,1579) = 2922.083$	$p = .000$	.649
	7th vs. 8th	$F(1,1579) = 145.322$	$p = .000$	.480
Free/Reduced	6th vs. later	$F(1,1086) = 1429.179$	$p = .000$	.568
	7th vs. 8th	$F(1,1086) = 654.937$	$p = .000$	.376

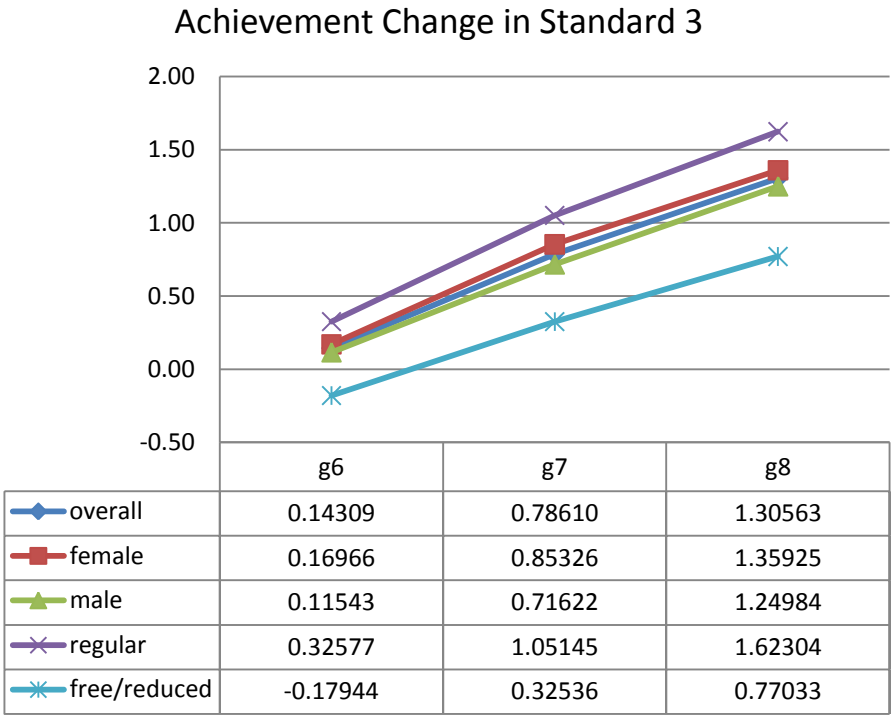


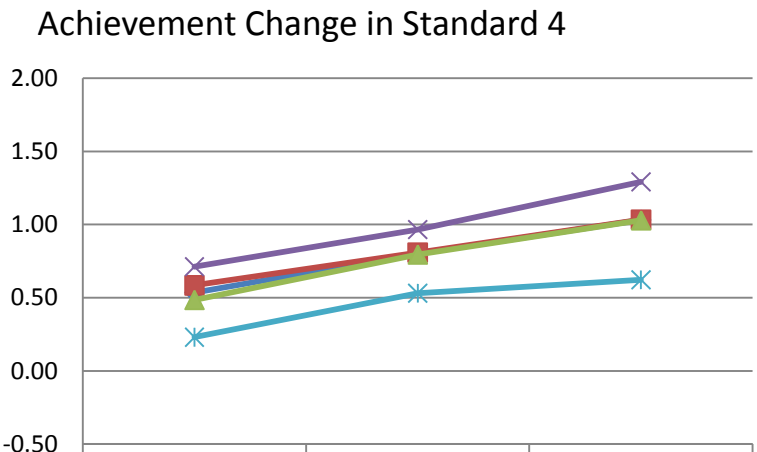
Figure 7.5 Achievement Change from 6th to 8th Grade in Standard 3 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.4 Standard 4

The two Helmert contrasts (6^{th} versus later and 7^{th} versus 8^{th}) in Standard 4 (Data) were significant for the overall data, F (1,2666) = 690.86, p = .000, η_p^2 = .206 and F (1,2666) = 331.50, p = .000, η_p^2 = .111, respectively. There was no difference for gender and the SES groups in the achievement changes over the three grades as shown in Table 7.11. As in the other standards, the achievement levels of the overall data, females and males were very similar in the three grades while the regular price lunch group was the highest and the free/reduced price lunch group was the lowest in those values as shown in Figure 7.6.

Table 7.11 *Significance of the Mathematical Achievement Change from 6^{th} to 8^{th} grade in Standard 4*

Group	*Contrast*	*F*	*p*	η_p^2
Overall	6^{th} vs. later	$F(1,2666) = 690.862$	$p = .000$	.206
	7^{th} vs. 8^{th}	$F(1,2666) = 331.496$	$p = .000$	.111
Female	6^{th} vs. later	$F(1,1359) = 273.103$	$p = .000$	.167
	7^{th} vs. 8^{th}	$F(1,1359) = 176.391$	$p = .000$	.115
Male	6^{th} vs. later	$F(1,1306) = 431.514$	$p = .000$	.248
	7^{th} vs. 8^{th}	$F(1,1306) = 156.131$	$p = .000$	.107
Regular	6^{th} vs. later	$F(1,1579) = 455.048$	$p = .000$	.224
	7^{th} vs. 8^{th}	$F(1,1579) = 386.710$	$p = .000$	.197
Free/Reduced	6^{th} vs. later	$F(1,1086) = 238.110$	$p = .000$	.180
	7^{th} vs. 8^{th}	$F(1,1086) = 22.856$	$p = .000$	.021



	g6	g7	g8
overall	0.53600	0.80199	1.03136
female	0.58462	0.80792	1.03336
male	0.48557	0.79583	1.02928
regular	0.71191	0.96563	1.29183
free/reduced	0.23036	0.53016	0.62315

Figure 7.6 Achievement Change from 6th to 8th Grade in Standard 4 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.5 Benchmark 1.1

The achievement level in Benchmark 1.1 (number sense) was significantly improved from 6th to 7th grade as well as from 7th to 8th grade, F (1,2666) = 3433.853, p = .000, η_p^2 = .563 for 6th grade versus later and F (1,2666) = 121.283, p = .000, η_p^2 = .044 for 7th versus 8th grade. No gender nor SES difference was found in the achievement changes over the three grades. Table 7.12 and Figure 7.7 present the significance test results of the two contrasts and the achievement changes for the five groups.

Table 7.12 *Significance of the Mathematical Achievement Change from 6th to 8th grade in Benchmark 1.1*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6th vs. later	$F(1,2666) = 3433.853$	$p = .000$	.563
	7th vs. 8th	$F(1,2666) = 121.283$	$p = .000$	.044
Female	6th vs. later	$F(1,1359) = 1772.977$	$p = .000$	.566
	7th vs. 8th	$F(1,1359) = 44.421$	$p = .000$	.032
Male	6th vs. later	$F(1,1306) = 1665.196$	$p = .000$	.560
	7th vs. 8th	$F(1,1306) = 82.057$	$p = .000$	.059
Regular	6th vs. later	$F(1,1579) = 2714.309$	$p = .000$	.631
	7th vs. 8th	$F(1,1579) = 74.639$	$p = .000$	.045
Free/Reduced	6th vs. later	$F(1,1086) = 935.144$	$p = .000$	.463
	7th vs. 8th	$F(1,1086) = 47.483$	$p = .000$	.042

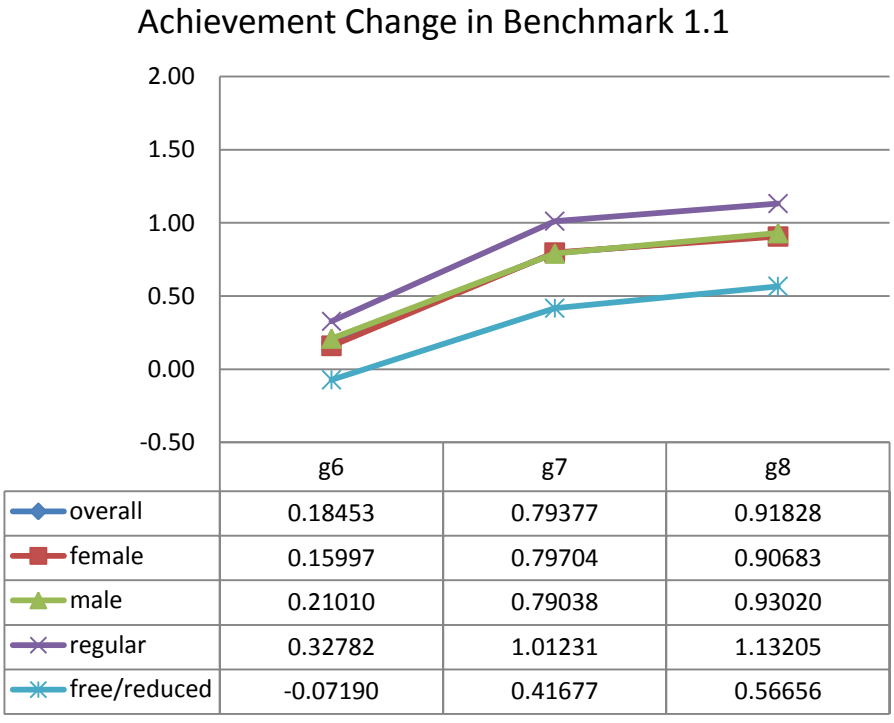


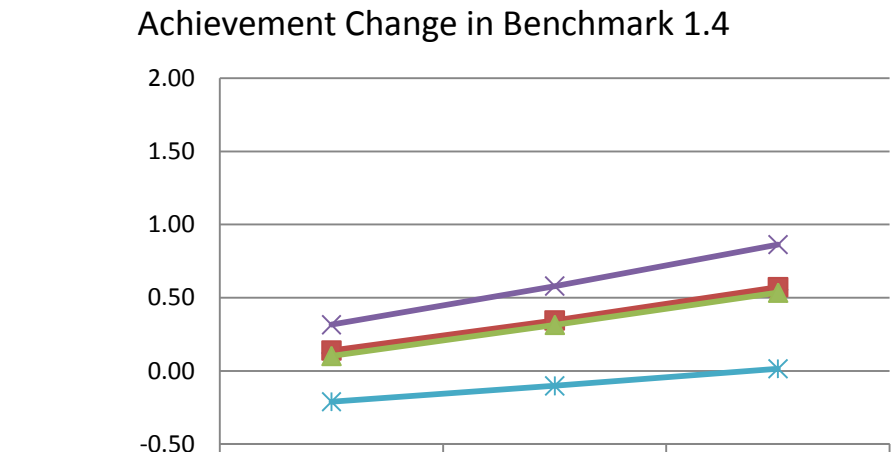
Figure 7.7 Achievement Change from 6th to 8th Grade in Benchmark 1.1 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.6 Benchmark 1.4

The two contrasts results in Benchmark 1.4 (computation) were presented in Table 7.13 and Figure 7.7 for each group. The achievement changes over the three grades in Benchmark 1.4 showed a clear linear trend for all the groups. For the overall data, the two contrasts were significant: (1) 6th grade versus later [F (1,2666) = 1166.96, p = .000, η_p^2 = .304] and (2) 7th versus 8th grade [F (1,2666) = 573.16, p = .000, η_p^2 = .177]. Both contrasts were also significant for all the four subgroups as presented in Table 7.13. Therefore, it can be concluded that there was neither gender nor SES difference in the pattern of the achievement changes over the three grades as shown in Figure 7.8.

Table 7.13 *Significance of the Mathematical Achievement Change from 6th to 8th grade in Benchmark 1.4*

Group	*Contrast*	*F*	*p*	η_p^2
Overall	6th vs. later	F(1,2666) = 1166.963	p = .000	.304
	7th vs. 8th	F(1,2666) = 573.158	p = .000	.177
Female	6th vs. later	F(1,1359) = 603.506	p = .000	.308
	7th vs. 8th	F(1,1359) = 294.409	p = .000	.178
Male	6th vs. later	F(1,1306) = 563.580	p = .000	.301
	7th vs. 8th	F(1,1306) = 278.586	p = .000	.176
Regular	6th vs. later	F(1,1579) = 1121.829	p = .000	.415
	7th vs. 8th	F(1,1579) = 539.412	p = .000	.255
Free/Reduced	6th vs. later	F(1,1086) = 196.805	p = .000	.153
	7th vs. 8th	F(1,1086) = 93.738	p = .000	.079



	g6	g7	g8
overall	0.12129	0.33010	0.55385
female	0.13992	0.34492	0.57184
male	0.10191	0.31468	0.53512
regular	0.31548	0.57982	0.86360
free/reduced	-0.20983	-0.10190	0.01439

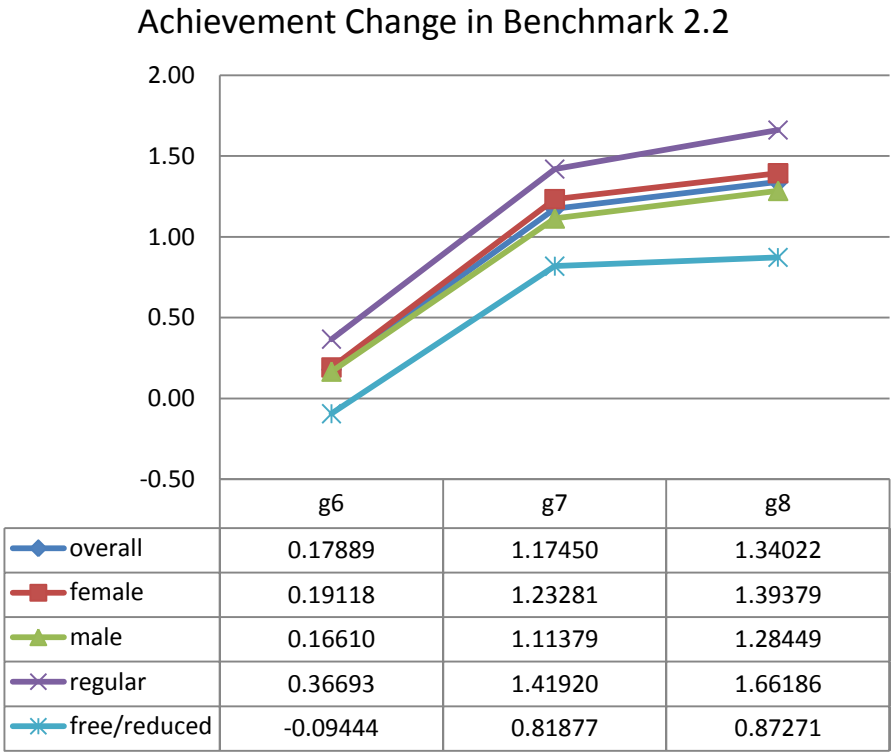
Figure 7.8 Achievement Change from 6th to 8th Grade in Benchmark 1.4 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.7 Benchmark 2.2

In Benchmark 2.2 (variables, equations, and inequalities), the two contrasts (6th versus later and 7th versus 8th) were significant for the overall data, F (1,2666) = 5773.58, p = .000, η_p^2 = .684 and F (1,2666) = 231.25, p = .000, η_p^2 = .08, respectively, as well as for all the four subgroups regarding gender and SES. However, the second contrast (7th versus 8th grade) showed a relatively small effect size (η_p^2 = .005) for the free/reduced price lunch group. The contrasts result for all the five categories of data were shown in 7.14 and Figure 7.9.

Table 7.14 *Significance of the Mathematical Achievement Change from 6th to 8th grade in Benchmark 2.2*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6th vs. later	$F(1,2666) = 5773.576$	$p = .000$	.684
	7th vs. 8th	$F(1,2666) = 231.252$	$p = .000$	.080
Female	6th vs. later	$F(1,1359) = 3341.086$	$p = .000$	.711
	7th vs. 8th	$F(1,1359) = 109.514$	$p = .000$	.075
Male	6th vs. later	$F(1,1306) = 2493.565$	$p = .000$	.656
	7th vs. 8th	$F(1,1306) = 122.140$	$p = .000$	.086
Regular	6th vs. later	$F(1,1579) = 4281.660$	$p = .000$	.731
	7th vs. 8th	$F(1,1579) = 285.318$	$p = .000$	.153
Free/Reduced	6th vs. later	$F(1,1086) = 1754.550$	$p = .000$	.618
	7th vs. 8th	$F(1,1086) = 11.193$	$p = .001$	.010



	g6	g7	g8
overall	0.17889	1.17450	1.34022
female	0.19118	1.23281	1.39379
male	0.16610	1.11379	1.28449
regular	0.36693	1.41920	1.66186
free/reduced	-0.09444	0.81877	0.87271

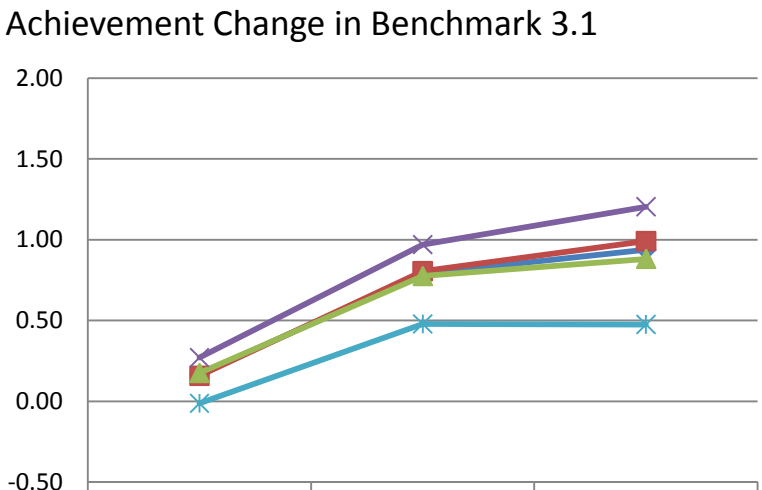
Figure 7.9 Achievement Change from 6th to 8th Grade in Benchmark 2.2 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.8 Benchmark 3.1

For all the groups except the free/reduced price lunch group, the two contrasts (6^{th} versus later and 7^{th} versus 8^{th}) in Benchmark 3.1 (geometric figures and their properties) were significant as shown in Table 7.15. For the free/reduced price lunch group, the achievement levels between 7^{th} and 8^{th} grades showed statistically no difference, $F(1,1086) = 1.20$, $p = .274$, and , $\eta_p^2 = .001$. As shown in Figure 7.10, there was no improvement in the achievement in Benchmark 3.1 at 8^{th} grade for this group. This finding was inconsistent with the result of Standard 2 in which all the groups including the free/reduced price lunch group showed a clear linear trend in the achievement changes over the three grades. It was speculated that the relatively high difficulty of Benchmark 3.1 as shown in Figure 6.4 (average p-values in 8^{th} grade) may have caused the insignificant change at 8^{th} grade for the low SES group.

Table 7.15 *Significance of the Mathematical Achievement Change from 6^{th} to 8^{th} grade about Benchmark 3.1*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6^{th} vs. later	$F(1,2666) = 5773.576$	$p = .000$	.684
	7^{th} vs. 8^{th}	$F(1,2666) = 231.252$	$p = .000$	.080
Female	6^{th} vs. later	$F(1,1359) = 2125.619$	$p = .000$	.610
	7^{th} vs. 8^{th}	$F(1,1359) = 133.855$	$p = .000$	.090
Male	6^{th} vs. later	$F(1,1306) = 1511.879$	$p = .000$	.537
	7^{th} vs. 8^{th}	$F(1,1306) = 37.834$	$p = .000$	.028
Regular	6^{th} vs. later	$F(1,1579) = 3177.146$	$p = .000$	.668
	7^{th} vs. 8^{th}	$F(1,1579) = 257.863$	$p = .000$	.140
Free/Reduced	6^{th} vs. later	$F(1,1086) = 844.518$	$p = .000$	.437
	7^{th} vs. 8^{th}	$F(1,1086) = 1.198$	$\mathbf{p = .274}$	.001



	g6	g7	g8
overall	0.16721	0.79050	0.93779
female	0.15859	0.80497	0.99164
male	0.17617	0.77544	0.88176
regular	0.27152	0.97039	1.20455
free/reduced	-0.01207	0.47862	0.47580

Figure 7.10 Achievement Change from 6th to 8th Grade in Benchmark 3.1for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.9 Benchmark 3.4

Benchmark 3.4 (Geometry from an Algebraic Perspective) was measured only in 6th and 8th grade. The achievement change from 6th to 8th grade was very significant for all the groups as shown in Table 7.16 and Figure 7.11. The result was consistent with the result of Standard 3 where the achievement changes over the grades were significant for all the groups. There was neither gender nor SES difference for the achievement change. As in the other contents, the regular price lunch group showed the highest achievement level while the free/reduced price lunch group had the lowest level.

Table 7.16 *Significance of the Mathematical Achievement Change from 6th to 8th grade about Benchmark 3.4*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6th vs. 8th	*F*(1,2666) = 2231.618	*p* = .000	.456
Female	6th vs. 8th	*F*(1,1359) = 1260.865	*p* = .000	.481
Male	6th vs. 8th	*F*(1,1306) = 982.237	*p* = .000	.429
Regular	6th vs. 8th	*F*(1,1579) = 2128.691	*p* = .000	.574
Free/Reduced	6th vs. 8th	*F*(1,1086) = 235.859	*p* = .000	.178

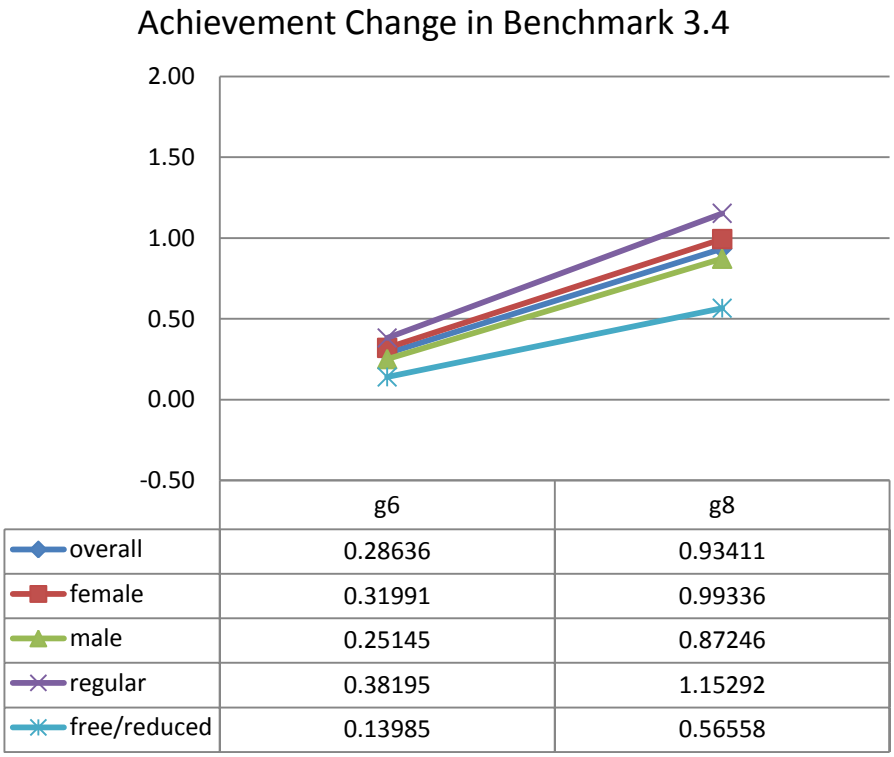


Figure 7.11 Achievement Change from 6th to 8th Grade in Benchmark 3.4 for Overall, Female, Male, Regular price lunch, and Free/Reduced price lunch Groups

7.3.2.10 Benchmark 4.1

Like Benchmark 3.4, Benchmark 4.1 (probability) was also measured only in 6th and 8th grade. Its achievement change from 6th to 8th grade was very significant for all the groups as shown in Table 7.17 and Figure 7.12. The result was consistent with the result

of Standard 4 in which the achievement changes over the three grades were significant and no group difference was found in the achievement change.

Table 7.17 *Significance of the Mathematical Achievement Change from 6th to 8th grade about Benchmark 4.1*

Group	*Contrast*	*F*	*P*	η_p^2
Overall	6th vs. 8th	*F*(1,2666) = 3530.239	*p* = .000	.570
Female	6th vs. 8th	*F*(1,1359) = 2014.928	*p* = .000	.597
Male	6th vs. 8th	*F*(1,1306) = 1540.814	*p* = .000	.541
Regular	6th vs. 8th	*F*(1,1579) = 2992.657	*p* = .000	.656
Free/Reduced	6th vs. 8th	*F*(1,1086) = 110.079	*p* = .000	.092

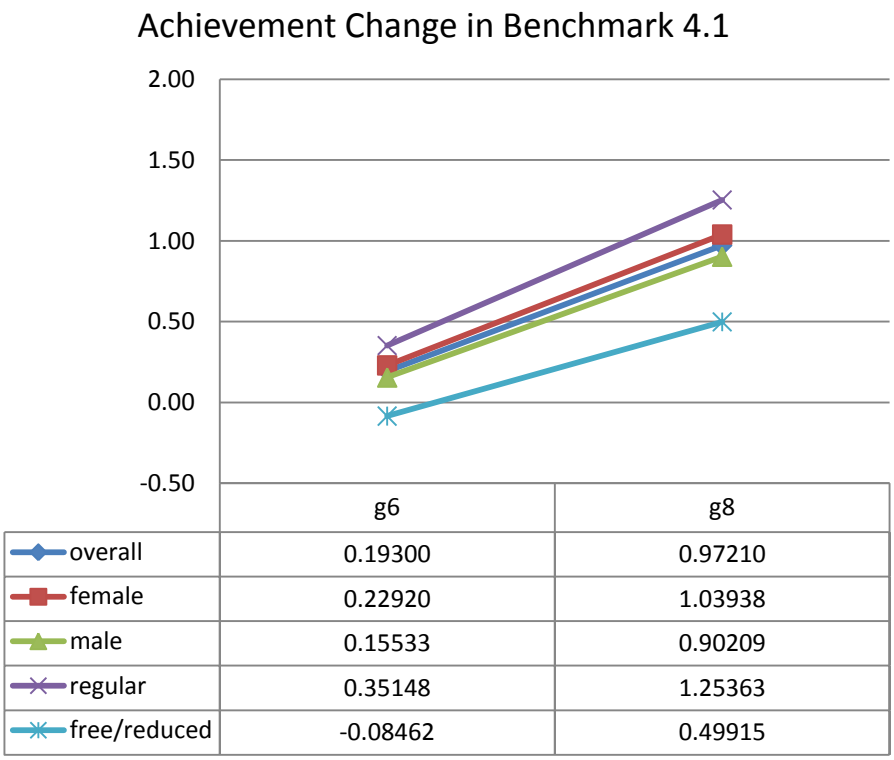


Figure 7.12 Achievement Change from 6th to 8th Grade in Benchmark 4.1 for Overall, Female, Regular price lunch, and Free/Reduced price lunch Groups

7.4 Study 2 Summary and Discussion

First, there were two contradictory hypotheses regarding the gender difference in mathematical achievement:

Hypothesis 6-a. There is no significant gender difference in mathematical achievement between female and male.

Hypothesis 6-b. Males outperform females in the mathematical achievement test.

From the competency level estimates at the three occasions (6^{th}, 7^{th}, and 8^{th} grade) in the FICFA models in Study 2.1, a gender difference was found in some mathematical contents. In Standard 4 (Data), females outperformed males at 6^{th} grade although no gender gap was found at 7^{th} and 8^{th} grade. On the other hand, in four out of the six benchmarks, such as Benchmarks 1.1 (number sense), Benchmark 1.4 (computation), Benchmark 2.2 (variables, equations, and inequalities), and Benchmark 3.1 (geometric figures and their properties), males outperformed females in 6^{th} grade; however, in 7^{th} and 8^{th} grade, no more gender difference was found except Benchmark 3.1. For Benchmark 3.1, interestingly, females surpassed males in 7^{th} and 8^{th} grade. Consistently with the result in Benchmark 3.1, in Standard 3 (Geometry), females outperformed males in 7^{th} and 8^{th} grade. Additionally, males outperformed females in Standard 2 (Algebra) at 8^{th} grade, but there was no gender gap in 6^{th} and 7^{th} grade. Therefore, hypothesis 6-a was generally supported because no gender difference was found in most mathematical contents in 7^{th} and 8^{th} grade. However, hypothesis 6-b was not supported well by the findings because boys and girls outperformed each other in some content areas; boys seemed to outperform in numbers/computation and algebra while girls seemed to do

better in data and geometry in general. Again, most gender gaps became less significant in 7th and 8th grade.

Second, one hypothesis exists regarding the SES factor on mathematical achievement:

*Hypothesis 7.*There is a significant SES difference in mathematical achievement.

In the competency level estimates by the FICFA models, a significant difference between the two SES groups was found in all the contents and occasions; the regular price lunch group outperformed the free/reduced price lunch group (low SES). The finding strongly supported Hypothesis 7 as well as the previous study results that low SES was linked to poor mathematical achievement. According to National Mathematics Advisory Panel (2008), SES is a potential source of the mathematical achievement gap. The conventional explanation for poor mathematical achievement for low SES students was the home environment factor such as quality of child-parent relationship, inadequate social experiences and learning opportunities (Crane, 1996; NMAP, 2008).

Third, like the FICFA model, it was explored how to apply a bifactor model to a longitudinal data using the three time factors (6th, 7th, and 8th grade). It was found that the loadings of all the items on the general factor (overall competency level of the content) were high, which indicated that all the test items well represented the general factor. However, the loadings on the three time factors showed different patterns in size depending on the mathematical contents. In some contents (Standard 1, 2, 4 and Benchmarks 2.2, 3.4, 4.1), 6th grade had higher factor loadings than 7th and 8th grades, which may suggest that these contents are more specific to 6th grade than the other grades. Also, the finding supports the previous theory that children develop their central

conceptual structure of number by ten years old (Case, 1985). On the other hand, in Standard 3 (Geometry), the time factor loadings were highest in 8th grade, which may imply that this content is more specific to 8th grade than the other grades, thus, the correlations between the Geometry items were largely accounted for by the 8th grade-factor. Interestingly, the finding supported the previous study result that the success in geometry was highly associated with logical reasoning and other mathematical skills (Tatsuoka et al., 2004) which can be obtained mostly in formal operation stage (7th or 8th grade). In conclusion, the different patterns of the three time factor loadings in each mathematical content may reflect the distinctiveness of the content regarding which grade level have bigger impact on the overall competency.

Finally, there was one primary hypothesis regarding the mathematical achievement changes from 6th to 8th grade:

Hypothesis 8. There is a significant incremental difference in mathematical achievement between 6th grade and the next two grade levels (7th and 8th); a quadratic trend of growth exists, such that the change from 6th to 7th grade is greater than from 7th to 8th grade.

MRMLC estimates showed a significant achievement change between 6th and the later two grades (7th and 8th) in all the contents (four standards and six benchmarks) regardless of gender and SES. Between 7th and 8th grade, a significant change was also found in all the contents for both genders and the regular price lunch group. However, for the free/reduced price lunch group (low SES), the change was not significant in some contents. Table 7.18 provides the summary of the mathematical achievement change over the three grades for each of the contents and the groups (gender and SES). As shown in

Table 7.18, the low SES group of students failed to improve their achievement at 8th grade in Standard 1 (Number and Computation) and Benchmark 3.1 (geometric figures and their properties).

Table 7.18. *Summary of the Mathematical Achievement Change from 6th to 8th grade for Each Group*

Content	*Overall*	*Female*	*Male*	*Regular price lunch*	*Free/ Reduced*
Standards 1	6<7<8	6<7<8	6<7<8	6<7<8	**6<7=8**
Standards 2	6<7<8	6<7<8	6<7<8	6<7<8	6<7<8
Standards 3	6<7<8	6<7<8	6<7<8	6<7<8	6<7<8
Standards 4	6<7<8	6<7<8	6<7<8	6<7<8	6<7,8
Benchmark 1.1	6<7<8	6<7<8	6<7<8	6<7<8	6<7<8
Benchmark 1.4	6<7<8	6<7<8	6<7<8	6<7<8	6<7<8
Benchmark 2.2	6<7<8	6<7<8	6<7<8	6<7<8	6<7<8
Benchmark 3.1	6<7<8	6<7<8	6<7<8	6<7<8	**6<7=8**
Benchmark 3.4	6<8	6<8	6<8	6<8	6<8
Benchmark 4.1	6<8	6<8	6<8	6<8	6<8

NOTE: < (significant increase), = (no significant change) at $\alpha = .01$

The main focus of this study was on examining how middle school students' mathematical achievement changes from 6th grade to 8th grade. The largest growth occurred from Grade 6 to Grade 7. The specific pattern of growth varied substantially by the socio-economic status of the student but few differences emerged by gender. Therefore, the result supported well Hypothesis 8.

CHAPTER 8

CONCLUSIONS

8.1 Summary and Findings

The importance of mathematics will continue to increase as development of more advanced technology is needed in the future. Mathematical achievement tests are universally applied throughout schooling in the U.S. to assess yearly progress. The middle school years (e.g., Grade 6-Grade 8) are especially crucial to success in mathematics because students must acquire the skills needed in Algebra and higher levels of mathematics (NMAP, 2008). The middle school years are also important developmentally because complex reasoning also emerges based on the developmental perspectives on cognition (e.g., Piaget, Vygotsky). Piaget insisted that children cannot learn from an experience until they have begun the transition into a stage that allows them to deal with and conceptualize that experience appropriately. Vygotsky's concept was that children can be improved when they have the assistance of the more advanced and competent people than themselves.

The purpose of the current study was to measure and interpret the mathematical achievement and the growth during the middle school years. According to many perspectives, the best design for studying the achievement change is a longitudinal study of representative samples of children. Until recently, however, inferences from such data were limited by the previous psychometric methods that were impractical to apply to large numbers of items and examinees. Therefore, some very recent advances in item response theory such as MRMLC with the IRTPRO program were applied to provide

inference about growth. For the current study, item responses to mathematical achievement tests administered during the middle school years were available for a randomly selected sample of 2,667 students in a Midwestern state.

This study had two parts: Study 1(analysis of item difficulty and factor structure) and study 2 (measuring mathematical achievement change). Study 1 consisted of Study 1-1(item difficulty), Study 1-2 (full information confirmatory factor analysis), and Study 1-3 (bifactor analysis by content). Study 2 (measuring mathematical achievement change) included Study 2-1(FICFA and bifactor analysis and by time factors) and Study 2-2 (MRMLC analysis).

Study 1 was an initial step to examine the general properties of the mathematical achievement test such as item difficulties and the factor structure. Study 1-1 was conducted to estimate item difficulty of the four standards and the six benchmarks, testing hypotheses #1 and #2:

Hypothesis 1. Difficulties of the four standards are in the following order: Standard 3 (Geometry) > Standard 2 (Algebra) > Standard 4 (Data) > Standard 1 (Number and computation).

Hypothesis 1 was not supported because the standards' difficulties were not significantly different within grades.

Hypothesis 2. Difficulties of the six specific benchmarks are in the following order: Benchmark 3.4 (geometry from an algebraic perspective) > Benchmark 3.1 (geometric figures and their properties) > Benchmark 2.2 (variables/equations and inequalities) > Benchmark 4.1 (probability) > Benchmark 1.4 (computation) > Benchmark 1.1 (number sense).

Hypothesis 2 was not supported because the benchmarks' difficulties were not significantly different within grades either.

In Study 1-2, the FICFA was conducted to test hypotheses #3 and #4:

Hypothesis 3. The four factor structure will be identified for each grade.

Hypothesis 3 was well supported because chi-square tests and fit indices indicated the existence of the four-factor structure by the four standards in the mathematical achievement test in all the three grades.

Hypothesis 4. The FICFA model with split loadings fits the data better than the FICFA model without split loadings.

Hypothesis 4 was well supported because the split loadings (involvement of items in multiple standards) were very significant in all the grades.

In Study 1-3, bifactor analysis was conducted to test hypothesis #5:

Hypothesis 5. The bifactor structure which includes a general factor (mathematical competency) and content-specific group factors (four standards or six benchmarks) exists across the grades.

Hypothesis 5 was supported well by the significant chi-square changes and high factor loadings of all the items on the general factor, which strongly supported the existence of a general factor in the mathematical achievement test in all the grades.

Study 2 was a primary study to measure and interpret the mathematical achievement and the growth during the middle school years. Growth in mathematical achievement was studied in the ten areas covered by the four standards (Number, Algebra, Geometry, and Data) and the six specific benchmarks. Differences in growth were also studied in two areas of individual differences, gender and SES background, that have

often been found important in careers that involve mathematics. In Study 2-1, FICFA and bifactor analysis using three time factors (Grades 6, 7, and 8) were conducted, testing hypotheses #6 and #7. Two contradictory hypotheses regarding the gender difference were tested as follows:

Hypothesis 6-a. There is no significant gender difference in mathematical Achievement.

Hypothesis 6-b. Males outperform females in the mathematical achievement test. Hypothesis 6-a was generally supported because no gender difference was found in most mathematical contents, especially, in 7^{th} and 8^{th} grade. However, hypothesis 6-b was not supported by the findings because boys and girls outperformed each other in some content areas; boys seemed to outperform in numbers/computation and algebra while girls seemed to do better in data and geometry in general. Again, most gender gaps became less significant in 7^{th} and 8^{th} grade.

*Hypothesis 7.*There is a significant SES difference in mathematical achievement. The finding strongly supported Hypothesis 7. A significant difference between the two SES groups was found in all the contents and all the three grades; students of the regular price lunch group performed significantly better than students of the free/reduced price lunch group. The finding also supported the results of many previous study results that low SES was linked to poor mathematical achievement.

In Study 2-2, multidimensional Rasch model for learning and change (MRMLC) analysis was conducted to measure students' mathematical achievement levels and the growth over the three middle school years:

Hypothesis 8. There is a significant incremental difference in mathematical achievement between 6th grade and the next two grade levels (7th and 8th); a quadratic trend of growth exists, such that the change from 6th to 7th grade is greater than from 7th to 8th grade.

The result in this study partly supported Hypothesis 8 because a significant achievement incremental change was found between 6th and 7th in all the mathematical contents (four standards and six benchmarks) regardless of gender and SES (free/reduced price lunch). However, the achievement change between 7th and 8th grade was also significant for all the groups except the low SES students in two areas (Standard 1 and Benchmark 3.1).

8.2 Discussion

First, Piaget's perspective of developmental stages was partly supported by the significant increase from 6th to 7th grade; some contents (Standard 1 and Benchmark 3.1) showed even decrease from 7th to 8th grade. As mentioned in Chapter 7, it was assumed that 6th graders were in concrete operational stage (age 6 or 7 to age 11 or 12) and 7th and 8th graders were in formal operational stage (age 11 or 12 and up) in this study. The largest growth from Grade 6 to Grade 7 in many content areas (i.e., Standards 1 and 2, Benchmarks 1.1, 2.2, and 3.1) seemed to reflect well the change from concrete operational stage to formal operational stage. However, a strong linear trend in the growth from Grade 6 through 8 was found in a few content areas (Standards 3 and 4, Benchmarks 1.4). The four benchmarks (1.1, 1.4, 2.2, and 3.1) were very useful for analyzing the shift in the middle school years (Grade 6 – Grade 8) because they were

applied for all the three grades. Notice that the requirement (indicators) of a benchmark becomes more cognitively complex for higher grades. For example, in regard to Benchmark 1.1 (number sense), student should be able to demonstrate number sense for rational numbers (at 6^{th} grade), irrational number pi (at 7^{th} grade), and real numbers (at 8^{th} grade) and simple algebraic expressions in one variable in a variety of situations. As elaborated in Chapter 2, when children enter formal operations stage, they become able to think and reason about abstract concepts, hypothetical ideas, and contrary-to-fact statements (Ormrod, 2008). Irrational numbers and real numbers may be abstract concepts in mathematics compared with rational numbers. Therefore, it appears that the indicators of the benchmarks were developed based on cognitive developmental stages of students and the result of study 2-2 supports the Piaget's developmental stages (concrete operational and formal operational stages) in three of the four benchmarks (1.1, 2.2, and 3.1).

Second, based on Vygotsky's theory, children can accomplish more difficult tasks when they have the assistance of people more advanced and competent than themselves. Also, consistent association has been found between SES and mathematical achievement in many studies over the years. The result of the current study support Vygotsky's view by the finding that the low SES group (free/reduced lunch students) underperformed the higher SES group (regular lunch students) in all the mathematical contents and the growth from grade 7 to grade 8 was not significant only for the low SES group in some contents (Standard 1 and Benchmark 3.1). As a result, it appeared that SES was a primary source of poor mathematical achievement in this study. If the low SES students had more

advanced metacognitive knowledge and skills about mathematics based on Vygotsky's social-cultural theories, their mathematical achievement could be significantly improved. Therefore, the result of this study can be better explained by both theories: Piaget's developmental stages and Vygotsky's social-cultural theory.

8.3 Limitations and Future Study

For the SES group classification, only school lunch program information was used in this study. However, SES has many definitions including parental education, poverty level, parental income, or a composite index (NMAP, 2008). In some studies, it was suggested that home environment factors are more important than SES such as the number of stimulating toys, quality of child- parent relationship (Crane, 1996), inadequate social experiences and learning opportunities in academic achievement (NMAP, 2008). Thus, more factors need to be included to explain students' mathematical achievement in future study.

Also, it would be interesting to examine which factor is stronger factor for the achievement differences between race and SES. Of the total 2667 students, 1580 (59.2%) were the regular price-lunch group and 1087 (41.7%) were the free/ reduced price-lunch group. However, of the regular price-lunch group, 1325 (84%) were Whites, 79 (5%) were Hispanics, and 66 (4%) were African Americans. In the free and reduced price lunch group, 491 (45%) were Whites, 290 (27%) were Hispanics, and 205 (19%) were African Americans. Therefore, it was speculated that race may affect the achievement in each SES group because the proportions of race differ in the two groups. A more balanced proportion in each SES group or factorial ANOVA would be helpful in

exploring the unique effects of SES and race as well as the interaction in the mathematical achievement and the change. In addition to, finding the interaction with gender and SES (or race) would be interesting research for future.

APPENDIX A

STANDARDS, BENCHMARKS, AND INDICATORS FOR MIDDLE SCHOOL MATHEMATICS (KANSAS STATE DEPARTMENT OF EDUCATION, 2003)

There are four levels of standard. First, Standard 1 is number and computation; students use numerical and computational concepts and procedures in a variety of situations. Second, Standard 2 is algebra; students use algebraic concepts and procedures in a variety of situations. Third, Standard 3 is geometry; students use geometric concepts and procedures in a variety of situations. Fourth, Standard 4 is data; students use concepts and procedures of data analysis in a variety of situations.

Standard 1 (Number and Computation)

For 6^{th} grade, benchmarks and their indicators are as follows:

- Benchmark 1.1 (Number sense): students demonstrate number sense for rational numbers and simple algebraic expressions in one variable in a variety of situations.
    - Indicator 1.1.1: Compares and orders - a) integers; b) fractions greater than or equal to zero; c) decimals greater than or equal to zero through thousandths place. Compare and order numbers to see which is larger or smaller for a set of numbers that include positive and negative numbers, fractions and decimals greater than 0.
    - Indicator 1.1.2: Knows and explains numerical relationships between percents, decimals, and fractions between 0 and 1. Know how to convert between percents, decimals, and fractions between 0 and 1 (25%=0.25=1/4)
- Benchmark 1.3 (Estimation): students use computational estimation with rational numbers and the irrational number pi in a variety of situations.
    - Indicator1.3.1: Estimates to check whether or not the result of a real-world problem using rational numbers and/or the irrational number pi is reasonable and makes predictions based on the information. In real life situations, it is a good idea to do estimation to check whether the exact answer is reasonable and to justify if it is or isn't a reasonable answer.
- Benchmark 1.4 (Computation): students model, perform, and explain computation with positive rational numbers and integers in a variety of situations.
    - Indicator 1.4.1: Performs and explains these computational procedures: a) divides which numbers through a 2-digit divisor and a 4-digit dividend and expresses the remainder as a whole number, fraction, or decimal. Use division of whole numbers greater than 0 by up 2-digit that may have remainders. If there is a remainder, express as a decimal or fraction.
    - Indicator 1.4.2: Generates and/or solves one- and two-step real-world problems with rational numbers using the computational procedures: b) addition, subtraction, multiplication, and division of decimals through hundredths place. Solve real-world problems by using one or two operations including addition, subtraction, multiplication, and/or division.

For 7th grade, benchmarks and their indicators are as follows:

- Benchmark 1.1 (Number sense): students demonstrate number sense for rational numbers, the irrational number pi, and simple algebraic expressions in one variable in a variety of situations.
    - Indicator 1.1.1: Generates and/or solves real-world problems using:a) equivalent representations of rational numbers and simple algebraic expressions: b) addition, subtraction, multiplication, and division of rational numbers with a special emphasis on fractions and expressing answers in simplest form. Students realize that there are a variety of ways to represent expressions such as 2× is the same as × + × or $.50 can be represented with two quarters ($.25 = $.25) or five dimes ($.10 + $.10 + $.10 + $.10 + $.10). Use equivalent representations for fractional operations such as 2/4 + 2/4 which equals 1 is the same as 1/2 + 1/2, or 2/4 × ¾ which equals 3/8 is the same as 1/2 × 3/4.
- Benchmark 1.4 (Computation): students model, perform, and explain computation with rational numbers, the irrational number pi, and first-degree algebraic expressions in one variable in a variety of situations.
    - Indicator 1.4.1: Performs and explains these computational procedures: a) adds and subtracts decimals from ten millions place through hundred thousandths place; b) multiplies and divides a four-digit number by a two-digit number using numbers from thousands place through thousandths place; c) multiplies and divides using numbers from thousands place through thousandths place by 10; 100; 1,000; 0.1; 0.01; 0.001; or signal-digit multiples of each; d) adds, subtracts, multiplies, and divides fractions and expresses answers in simplest form. Follows are the examples of the four types above.
        a) ten millions, millions, hundred thousands, ten thousands, thousands, hundreds, tens, ones, tenth, hundredth, thousandth, ten thousandth, hundred thousandth
        b) 1.698 ÷ 25 or 1.698 × 25
        c) 54.3 ÷ .002 or 54.3 × .002
        d) 2/3 ×5/8
    - Indicator 1.4.2: Finds percentages of rational numbers-Rational number is any number that can be written as a fraction. Percent is based on an amount out of 100.

For 8th grade, benchmarks and their indicators are as follows:

- Benchmark 1.1 (Number sense): students demonstrate number sense for real numbers and simple algebraic expressions in one variable in a variety of situations.
    - Indicator 1.1.1: Knows and explains what happens to the product or quotient when: a) a positive number is multiplied or divided by a rational number grater than zero and less than one; b) a positive number is multiplied or divided by a rational number greater than one; c) a nonzero real number is multiplied or divided by zero.When a positive number (such as 4) is multiplied by a number greater than

zero but less than 1 (such as 1/2), the result is smaller than the first number (4 × 1/2 = 2). When a positive number (such as 4) is divided by a number greater than zero but less than 1 (such as 1/2), the result is greater than the first number (4 ÷ 1/2 = 8). When a number other than 0 (such as -4) is multiplied by 0, the result is 0 (-4×0 = 0).

- Benchmark 1.2 (Number systems and their properties):students demonstrate an understanding of the real number system; recognizes, applies, and explains their properties; and extends these properties to algebraic expressions.
    - Indicator 1.2.1: Identifies all the subsets of the real number system [natural (counting) numbers, whole numbers, integers, rational numbers, irrational numbers] to which a given number belongs.Natural numbers are those numbers we count with; whole numbers are the counting numbers and zero; integers include zero, whole numbers and their opposites; rational numbers are those numbers that can be expressed as fractions; and irrational numbers are those numbers that cannot be expressed as fractions. Students need to know which set(s) of numbers to which a given number belongs
    - Indicator1.2.2: Generates and or/solves real-world problems with rational numbers using the concepts of these properties to explain reasoning: a) commutative, associative, distributive, and substitution properties; b) identity and inverse properties of addition and multiplication. Numbers can be added or multiplied in any order resulting with the same answer (commutative). When a series of numbers is added or multiplied, the order in which the values are added or multiplied doesn't affect the result (associative). When multiplying a number by the sum of numbers, you can multiply each of the numbers by the factor fist and then add (distributive). A number may be substituted for a variable or equivalent quantity (substitution). When 0 is added to another number it doesn't change the value of the number (identify for addition). When a number is multiplied by 1 it doesn't change the value of the number (identify for multiplication). A number plus its opposite is 0 (additive inverse). A number multiplied by its reciprocal is 1 (multiplicative inverse). It is important that students know the name of the property as used in the indicator.
- Benchmark 1.4 (Computation): students model, perform, and explain computation with rational numbers, the irrational number pi, and algebraic expressions in a variety of situations.
    - Indicator 1.4.1: Performs and explains these computational procedures with rational numbers: a) addition, subtraction, multiplication, and division of integers; b) order of operations (evaluates within grouping symbols, evaluates powers to the second or third power, multiplies or divides in order from left to right, then adds or subtracts in order from left to right).Compute with integers (positive and negative whole numbers and zero). Use order of operations when computing with rational numbers. (First work within grouping symbols; then find

powers; then perform multiplication/division left to right; then perform addition/subtraction left to right.)

- Indicator 1.4.2: Generates and/or solves one- and two-step real-world problems using computational procedures and mathematical concepts: a) rational numbers; b) the irrational number pi as an approximation; c) applications of percents.Solve real-world problems that involve one and two steps to solve using computation of addition, subtraction, multiplication, and/or subtraction of numbers that are positive and negative with decimals 9that repeat or terminate), fractions, or use pi (3.1415926…) and with percents.

Standard 2 (Algebra)

For 6th grade, benchmarks and their indicators are as follows:

- Benchmark 2.1 (Patterns): students recognize, describe, extend, develop, and explain the general rule of a pattern in variety of situations.
  - Indicator 2.1.1: States the rule to find the next number of a pattern with one operational change (addition, subtraction, multiplication, division) to move between consecutive terms. Find the rule 9addition, subtraction, multiplication, or division) for a set of numbers in a pattern in order to extend the pattern past the last term given.
- Benchmark 2.2 (Variables, Equations, and Inequalities): students use variables, symbols, positive rational numbers, and algebraic expressions in one variable to solve linear equations and inequalities in a variety of situations.
  - Indicator 2.2.1: Represents real-world problems using variables and symbols to: b) write and/or solve one-step equations (addition, subtraction, multiplication, and division).Write and solve one-step equations for real-world situations using addition, subtraction, multiplication and division.

For 7th grade, benchmarks and their indicators are as follows:

- Benchmark 2.1 (Patterns): students recognize, describe, extend, develop, and explain the general rule of a pattern in variety of situations.
  - Indicator 2.1.1: Identifies, states, and continues a pattern presented in various formats including numeric (list or table), algebraic (symbolic notation), visual (pictures, table, or graph), verbal (oral description), kinesthetic (action), and written using these attributes: a) counting numbers including perfect squares, cubes, and factors and multiples (number theory); b) positive rational numbers including arithmetic and geometric sequences (arithmetic: sequence of numbers in which the difference of two consecutive numbers is the same, geometric: a sequence of numbers in which each succeeding term is obtained by multiplying the preceding term by the same number).Identifies and extends a variety of patterns.
  - Indicator 2.1.2: States the rule to find the n^{th} term of a pattern with one operational change (addition or subtraction) between consecutive

terms. The n^{th} term is an arbitrary term is sequence or pattern of numbers which can be found by with the rule for the pattern.

- Benchmark 2.2 (Variables, Equations, and Inequalities): students use variables, symbols, rational numbers, and simple algebraic expressions in one variable to solve linear equations and inequalities in a variety of situations.
    - Indicator 2.2.1: Knows the mathematical relationship between ratios, proportions, and percents and how to solve for a missing term in a proportion with positive rational number solutions and monomials. Knows that ¾ (ratio) is equal to 75%. Knows that ¾ = 6/8 (a proportion), and are both equal to 75%
    - Indicator 2.2.2: Evaluates simple algebraic expressions using positive rational numbers. Replacing variables (what's unknown) with given numbers and finding the value.
    - Indicator 2.2.3: Represents real-world problems using variables and symbols to write linear expressions, one- or two- step equations.Write expressions (2.89×) and equations (2.89× = 12). An expression does not contain an equal sign while an equal does.

For 8^{th} grade, benchmarks and their indicators are as follows:

- Benchmark 2.2 (Variables, Equations, and Inequalities): students use variables, symbols, rational numbers, and algebraic expressions to solve linear equations and inequalities in a variety of situations.
    - Indicator 2.2.1: Solves: a) one- and two-step linear equations in one variable with rational number coefficients and constants intuitively and/or analytically.Find the solution to an equation. Students may choose to do this in any way that is successful and makes sense to them
    - Indicator 2.2.2: Represents real-world problems using: a) variables, symbols, expressions, one- or two- step equations with rational number coefficients and constants.Use variables, symbols, expressions, or equations to represent unknown quanities to represent real-world problems and solve using computation of addition, subtraction, multiplication, and/or subtraction of numbers that are positive and negative with decimals (that repeat or terminate), and fractions. A variable or symbol is used to represent an unknown and known numbers; the variable can be replaced for the number to solve.
- Benchmark 2.3 (Functions): students recognize, describ, and analyze constant, linear and nonlinear relationships in a variety of situations.
    - Indicator 2.3.1: Translates between the numerical, tabular, graphical, and symbolic representations of linear relationships with integer coefficients and constants.Change (translate) between numerical, tabular, graphical, and symbolic representations of a set of data that is linear.
- Benchmark 2.4 (Models): students generate and use mathematical models to represent and justify mathematical relationships found in a variety of situations.

- Indicator 2.4.1: Determines if a given graphical, algebraic, or geometric model is an accurate representation of a given real-world situation. Mathematical models are representations of some type of situation within a mathematical situation. They can be graphical (such as a picture representing 10,000 people in a report of population), algebraic (such as a formula to represent the area of a circle), or geometric (such as a geometric shape on a balance scale represents a given weight).

Standard 3 (Geometry)

For 6th grade, benchmarks and their indicators are as follows:

- Benchmark 3.1 (Geometric Figures and Their Properties): students recognize geometric figures and compares their properties in a variety of situations.
  - Indicator 3.1.1: Classifies: a) angles as right, obtuse, acute, or straight;
  - b) triangles as right, obtuse, acute, scalene, isosceles, or equilateral. Classify angles as right, obtuse, acute, or straight; triangles by their angles as right, obtuse, or acute; and triangles by their sides as scalene, isosceles, or equilateral. Angles that are right = 90˚, obtuse are > 90˚ but <180˚, acute are <90˚ but 0˚, and straight = 180˚. Triangles that are right have one right angle, acute have three angles <90˚, and obtuse have one angle > 90˚. Triangles that are scalene have no sides the same length, isosceles have two sides the same length, and equilateral have all three sides the same length.
- Benchmark 3.2 (Measurement and Estimation): students estimate, measure, and use measurement formulas in a variety of situation.
  - Indicator 3.2.1: Coverts: b) within the metric system using the prefixes: kilo, hector, deka, deci, centi, and mili. Covert within the metric system of measurement for length (meters), mass (grams), and volume (liters) for the prefixes kilo, hector, deka, deci, centi, and milli. Convert means to change from one measurement to another that is equivalent (in the customary measurement system 24 inches would equal 2 feet).
  - Indicator 3.2.2: Solves real-word problems by applying these measurement formulas: a) perimeter of polygons using the same unit of measurement; b) areas of squares, rectangles, and triangles using the same unit of measurement. Solve real-world problems using perimeter of a variety of shapes with the same unit of measure and area of squares, rectangles, and triangles with the same unit of measure. Perimeter is distance around a figure and area is the amount of surface covered inside the figure.
- Benchmark 3.3 (Transformational Geometry): students recognize and perform transformations on two- and three-dimensional geometric figures in a variety of situations.

    - Indicator 3.3.1: Identifies, describes, and performs one or two transformations (reflection, rotation, translation) on a two-dimensional figure. The student will understand, demonstrate, and explain one or two transformations (reflection/flip, rotation/turn, and/or translation/side) of an object. Flip is to flip over, turn is to rotate, and slide is to move across a flat surface.
- Benchmark 3.4 (Geometry From an Algebraic Perspective): students relate geometric concepts to a number line and a coordinate plane in a variety of situations.
    - Indicator 3.4.1: Uses all four quadrants of the coordinate plane: a) identify the ordered pairs of integer values on a given graph; b) plot the ordered pairs of integer values. Identify and plot points on a coordinate system divided into four areas (quadrants). The coordinate plane is divided into four sections by two number lines (labeled the horizontal and vertical axis with two letters such as x and y) that perpendicularly intersect to form the origin (starting point for graphing or identifying points already on the plane. An ordered pair such as (3, -2) means three positive units to the right of the origin and then 2 units down from the horizontal axis.

For 7th grade, benchmarks and their indicators are as follows:

- Benchmark 3.1: Geometric Figures and Their Properties - The student recognizes geometric figures and compares their properties in a variety of situations.
    - Indicator 3.1.1: Identifies angle and side properties of triangles and quadrilaterals: a) sum of the interior angles of any triangle is 180˚; b) sum of the interior angles of any quadrilateral is 360˚; c) parallelograms have opposite sides that are parallel and congruent; d) rectangles have angles 90˚ opposite sides are congruent; e) rhombi have all sides the same length, opposite angles are congruent; f) squares have angles of 90˚, all sides congruent; g) trapezoids have one pair of opposite sides parallel and the other pair of opposite sides are not parallel. Recognize that shapes have specific characteristics such as; triangle has three angles that add up to 180˚, parallelograms have opposite sides that are parallel and congruent (same size, same shape), rectangle have four right angles (90˚), rhombi have four congruent sides, sequres have four congruent angles and sides, trapezoids are quadrilaterals (four sided figures) that have exactly one pair of parallel sides.
- Benchmark 3.2 (Measurement and Estimation): students estimate, measure, and use measurement formulas in a variety of situation.
    - Indicator 3.2.1: Knows and uses perimeter and area formulas for circles, squares, rectangles, triangles, and parallelograms. Find perimeter (distance around the outside) and area (square units of space inside) of various shapes

    - Indicator 3.2.2: Uses given measurement formulas to find: a) surface area of cubes; b) volume of rectangular prisms. Find surface area (the area of all six sides of a three dimensional object) and the volume of rectangular prisms (the amount of space inside of a box).
    - Indicator 3.2.3: Solves real-world problems: c) finding perimeter and area of two-dimensional composite figures of squares, rectangles, and triangles. Find distance around (perimeter) and space inside of (area) of figures made with suares, rectangles and triangles
- Benchmark 3.3 (Transformational Geometry): students recognize and perform transformations on two- and three-dimensional geometric figures in a variety of situations.
    - Indicator 3.3.1: Determines the actual dimensions and/or measurements of a two-dimensional figure represented in a scale drawing. Determine actual measurement of a distance given the scale of a drawing (e/g/ 1 in = 20 miles on a map).

For 8^{th} grade, benchmarks and their indicators are as follows:

- Benchmark 3.1 (Geometric Figures and Their Properties): students recognize geometric figures and compare their properties in a variety of situations.
    - Indicator 3.1.1: Uses the Pythagorean theorem: a) determine if a triangle is a right triangle; b) find a missing side of a right triangle where the lengths of all three sides are whole numbers. The Pythagorean Theorem is a formula that states that if a triangle is a right triangle (has a 90˚ angle), then the sum of the squares of the two legs is equal to the square of the hypotenuse (the side opposite the right angle). Decide if a triangle is a right triangle. Find the missing side on a right triangle. The formula is $a^2 + b^{2 =} c^{2.}$
    - Indicator 2: Solves real-world problems by: a) using the properties of corresponding parts of similar and congruent figures. Solve real-world problems using knowledge that congruent figures are the same exact shape and size and their corresponding sides are the same length and their areas are same. Solve real-world problems using knowledge that similar figures are the same exact shape and their corresponding sides are proportional in length and their areas are proportional to the increase in the sides.
- Benchmark 3.4: Geometry From an Algebraic Perspective - The student uses an algebraic perspective to examine the geometry of two-dimensional figures in a variety of situations.
    - Indicator 1: Uses the coordinate plane to: a) list several ordered pairs on the graph of a line and find the slope of the line; b) recognize that ordered pairs that lie on the graph of an equation are solutions to that equation; c) recognize that points that do not lie on the graph of an equation are not solutions to that equation; d) determine the length of a side of a figure drawn on a coordinate plane with vertices having the same x- or y- coordinates. On the graph of a line, list points on the line. Recognize these points (ordered pairs) as solutions to the equation.

Find the slope (rate of change) of the graph. From a figure drawn on the graph, find the length of a side given two points.

Standard 4 (Data)

For 6th grade, benchmarks and their indicators are as follows:

- Benchmark 4.1 (Probability): students apply the concepts of probability to draw conclusions and to make predictions and decisions including the use of concrete objects in a variety of situations.
    - Indicator 4.1.1: List all possible outcomes of an experiment or simulation with a compound event composed of two independent events in a clear and organized way. List all the possible ways something can happen in an experiment with two events that happen together but are not related to one another. Example would b having three pairs of shoes (red, blue, and black) and four pair of socks (yellow, blue, black, and white) and picking one pair of shoes and a pair of socks at random.
    - Indicator 4.1.2: Represents the probability of a simple event in an experiment or simulation using fractions and decimals.The probability of an event happening in a random experiment is the ratio (fraction) of the number of successful outcomes as the numerator over the total number of outcomes as the denominator. Convert fraction to a decimal by dividing numerator by denominator to get a decimal. An example would be the probability of rolling an even number on a die would be 3/6 which can be simplified to 1/2 or 0.5 as a decimal (1 divided by 2).

For 7th grade, benchmarks and their indicators are as follows:

- Benchmark 4.2 (Statistics): students collect, organize, display, and explain numerical (rational numbers) and non-numerical data sets in a variety of situations with a special emphasis on measures of central tendency.
    - Indicator 4.2.1: Organizes, displays, and reads quantitative (numerical) and qualitative (non-numerical) data in a clear, organized, and accurate manner including a title, labels, categories, and rational number intervals using these data displays: a) frequency tables and line plots; b) bar, line, and circle graphs; c) Venn diagrams or other pictorial displays; d) charts and tables; e) stem-and –leaf plots (single); f) scatter plots; g) box-and whiskers plots. Read and make a) vertical and horizontal tables and charts, b) line, circle, and picture graphs, and c) scatter, stem-and- leaf, and box-and-whiskers plots
    - Indicator 4.2.2: Recognize and explains: a) misleading representations of data; b) the effects of scale or interval changes on graphs of data sets. Find misrepresentations of data that distorts the appearance of the data. How changing the vertical and horizontal parts (axis) of a graph can distort the appearance of the graph.

For 8th grade, benchmarks and their indicators are as follows:

- Benchmark 4.1 (Probability): students apply the concepts of probability to draw conclusions, generate convincing arguments, and make predictions and decisions including the use of concrete objects in a variety of situations.
    - Indicator 4.1.1: Finds the probability of a compound event composed of two independent events in an experiment, simulation, or situation. Find probability (likelihood of something happening) of two independent (not related or dependent) events happening concurrently (at the same time) or consecutively (one after the other).
    - Indicator 4.1.2: Makes predictions based on the theoretical probability a) a simple event in an experiment or simulation. Theoretical probability is the expected probability in an experiment. If a die is rolled, each number 1-6 has a 1/6 probability of being rolled. If a die is rolled 300 times it is expected that 6 would be rolled 50 times ($1/6 \times 300 = 50$ times).
- Benchmark 4.2 (Statistics): students collect, organize, display, and interpret numerical (rational) and non-numerical data sets in a variety of situations.
    - Indicator 4.2.1: Determines and explains the measure of central tendency (mode, median, mean) for a rational number data set. Calculate mean, median, and mode for a set of numbers. Mean is the sum of the values divided by the number of values, median is middle value when all valued are ordered, and mode is the value that appears the most.

APPENDIX B

FACTOR LOADINGS IN FULL INFORMATION CONFIRMATORY FACTOR ANALYSIS

Table B.1 *6th Grade: Factor Loadings of FICFA Model with Split Loadings*

	Item	F1	F2	F3	F4
1	s1_m6	0	0	0.38	0
2	s2_m6	0	0	0.36	0
3	s3_m6	0	0	0.4	0
4	s4_m6	0	0	0.33	0
5	s5_m6	0	0	0.41	0
6	s6_m6	0	0	0.39	0
7	s7_m6	0	0	0.26	0
8	s9_m6	0	0	0	0.53
9	s10_m6	0	0	0	0.66
10	s11_m6	0	0	0	0.47
11	s12_m6	0	0	0	0.49
12	s13_m6	0	0	0	0.68
13	s14_m6	0	0	0	0.65
14	s15_m6	0.57	0	0	0
15	s16_m6	0.42	0	0	0
16	s17_m6	0.36	0	0	0
17	s19_m6	0.56	0	0	0
18	s21_m6	0	0	0	0.56
19	s23_m6	0	0	0	0.49
20	s26_m6	0	0	0	0.67
21	s28_m6	0	0	0	0.47
22	s29_m6	0	0.55	0	0
23	s32_m6	0	0.63	0	0
24	s33_m6	0	0.61	0	0
25	s34_m6	0	0.38	0	0
26	s35_m6	0	0.66	0	0
27	s36_m6	0	0.55	0	0
28	s37_m6	0	0.47	0	0
29	s39_m6	0	0.55	0	0
30	s41_m6	0	0	0.58	0
31	s42_m6	0	0	0.44	0
32	s44_m6	0	0	0.5	0
33	s45_m6	0	0	0.39	0
34	s46_m6	0	0	0.46	0
35	s48_m6	0	0	0.4	0
36	s50_m6	0	0	0.38	0
37	s53_m6	0	0	0.55	0
38	s54_m6	0	0	0.64	0

Table B.1 (*continued*)

39	s55_m6	0	0	0.64	0
40	s56_m6	0	0	0.66	0
41	s57_m6	0	0	0.6	0
42	s58_m6	0	0	0.38	0
43	s59_m6	0.46	0	0	0
44	s62_m6	0.51	0	0	0
45	s63_m6	0.18	0	0	0
46	s65_m6	0.44	0	0	0.19
47	s66_m6	0.3	0	0	0.21
48	s67_m6	0.32	0	0	0.21
49	s71_m6	0	0	0.44	0
50	s72_m6	0	0	0.43	0
51	s73_m6	0	0	0.28	0
52	s74_m6	0	0	0.41	0
53	s75_m6	0.59	0	0	0
54	s76_m6	0.61	0	0	0
55	s79_m6	0.51	0	0	0
56	s80_m6	0.53	0	0	0
57	s81_m6	0.46	0	0	0
58	s83_m6	0.47	0	0	0
59	s85_m6	0.43	0	0	0
60	s86_m6	0.47	0	0	0

Table B.2 *7th Grade: Factor Loadings of FICFA Model with Split Loadings*

	Item	F1	F2	F3	F4
1	s1_m7	0.5	0.47	0	0
2	s2_m7	0.31	0.34	0	0
3	s4_m7	0.35	0.44	0	0
4	s5_m7	0.36	0.36	0	0
5	s6_m7	0	0.39	0	0
6	s7_m7	0.42	0.38	0	0
7	s8_m7	0	0.29	0	0
8	s9_m7	0	0.52	0	0
9	s13_m7	0	0.21	0.23	0
10	s16_m7	0	0	0.47	0
11	s19_m7	0	0.38	0.41	0
12	s20_m7	0	0.44	0.58	0
13	s21_m7	0	0.41	0.45	0
14	s22_m7	0	0.43	0.51	0
15	s24_m7	0.53	0	0	0
16	s25_m7	0.43	0	0	0
17	s26_m7	0.55	0	0	0
18	s27_m7	0.42	0	0	0
19	s29_m7	0	0.54	0	0
20	s31_m7	0	0.45	0	0
21	s32_m7	0.27	0.36	0	0
22	s33_m7	0	0.45	0	0
23	s34_m7	0.24	0.16	0	0
24	s35_m7	0.32	0.39	0	0
25	s37_m7	0.3	0.19	0	0
26	s39_m7	0.49	0	0	0
27	s42_m7	0.3	0.32	0	0
28	s44_m7	0.35	0.26	0	0
29	s46_m7	0.41	0.27	0	0
30	s47_m7	0.31	0.27	0	0
31	s48_m7	0	0	0.28	0
32	s49_m7	0.08	0.16	0.28	0
33	s50_m7	0	0	0.42	0
34	s52_m7	0	0	0.55	0
35	s53_m7	0	0.4	0.11	0
36	s54_m7	0	0.54	0	0
37	s55_m7	0	0.52	0	0
38	s57_m7	0	0.57	0	0

Table B.2 (*continued*)

39	s58_m7	0.43	0	0	0
40	s59_m7	0.41	0	0	0
41	s60_m7	0.43	0	0	0
42	s62_m7	0.53	0	0	0
43	s65_m7	0.43	0	0	0
44	s66_m7	0	0	0.42	0
45	s67_m7	0	0	0.51	0
46	s68_m7	0	0	0.08	0
47	s69_m7	0	0	0.42	0
48	s71_m7	0	0	0.44	0
49	s72_m7	0	0	0.38	0
50	s73_m7	0	0	0	0.52
51	s74_m7	0	0	0	0.38
52	s75_m7	0	0	0	0.46
53	s76_m7	0	0	0	0.33
54	s77_m7	0	0	0	0.4
55	s78_m7	0	0	0	0.46
56	s79_m7	0	0	0	0.47
57	s80_m7	0	0	0	0.67
58	s81_m7	0	0	0	0.67
59	s83_m7	0	0	0	0.37
60	s84_m7	0	0	0	0.43

Table B.3 *8th Grade: Factor Loadings of FICFA Model with Split Loadings*

	Item	F1	F2	F3	F4
1	s1_m8	0.48	0	0	0
2	s2_m8	0.48	0.24	0	0
3	s3_m8	0.53	0	0	0
4	s4_m8	0.65	0	0	0
5	s5_m8	0.52	0	0.05	0
6	s6_m8	0.36	0	0	0
7	s7_m8	0.51	0	0	0
8	s8_m8	0.4	0	0	0
9	s10_m8	0.41	0	0	0
10	s11_m8	0.27	0	0	0
11	s12_m8	0.33	0	0	0
12	s13_m8	0.25	0	0	0.47
13	s14_m8	0	0	0	0.49
14	s15_m8	0	0	0	0.6
15	s16_m8	0	0	0	0.51
16	s17_m8	0	0	0	0.47
17	s18_m8	0.28	0	0	0.3
18	s20_m8	0.45	0	0.67	0
19	s22_m8	0.48	0	0.63	0
20	s23_m8	0.42	0	0.44	0
21	s24_m8	0	0.32	0	0.16
22	s26_m8	0	0.39	0	0.33
23	s29_m8	0	0.39	0	0.35
24	s39_m8	0	0.5	0	0
25	s42_m8	0	0	0	0.47
26	s43_m8	0	0.24	0	0.43
27	s44_m8	0	0.12	0	0.22
28	s45_m8	0	0.35	0	0.34
29	s46_m8	0	0.34	0	0.53
30	s47_m8	0	0	0	0.54
31	s48_m8	0	0	0	0.39
32	s49_m8	0	0.45	-0.01	0.27
33	s51_m8	0	0.48	0.07	0
34	s54_m8	0.36	0.26	0	0
35	s55_m8	0.25	0.41	0	0
36	s58_m8	0.25	0.35	0.07	0
37	s60_m8	0	0	0	0.57
38	s63_m8	0.27	0	0	0.46

Table B.3 (*continued*)

39	s73_m8	0.34	0.44	0	0
40	s74_m8	0	0.55	0	0
41	s75_m8	0	0.32	0	0
42	s76_m8	0	0.46	0	0
43	s80_m8	0.45	0	0	0
44	s81_m8	0.51	0	0	0
45	s82_m8	0.37	0	0	0
46	s84_m8	0.32	0	0	0
47	s86_m8	0.31	0	0	0
48	s88_m8	0.17	0	0	0
49	s89_m8	0.31	0	0	0
50	s90_m8	0	0	0.39	0
51	s91_m8	0	0.44	0.35	0
52	s93_m8	0	0.49	0.28	0
53	s94_m8	0	0.39	0.31	0
54	s95_m8	0	0.32	0.24	0
55	s96_m8	0	0.42	0.25	0
56	s97_m8	0	0.38	0.29	0
57	s98_m8	0	0.51	0	0
58	s99_m8	0	0.23	0	0
59	s100_m8	0	0.38	0	0
60	s101_m8	0	0.45	0	0

APPENDIX C

FACTOR LOADINGS IN BIFACTOR MODLES BY CONTENT

Table C.1 *6th Grade: Factor Loadings of Bifactor Model for Four Standards*

	Item	g	F1	F2	F3	F4
1	s1_m6	0.53	0	0	0	0
2	s2_m6	0.53	0	0	0	0
3	s3_m6	0.56	0	0	0	0
4	s4_m6	0.53	0	0	0	0
5	s5_m6	0.56	0	0	0	0
6	s6_m6	0.58	0	0	0	0
7	s7_m6	0.34	0	0	0	0
8	s9_m6	0.52	0	0	0	0.34
9	s10_m6	0.64	0	0	0	0.46
10	s11_m6	0.46	0	0	0	0.29
11	s12_m6	0.51	0	0	0	0.28
12	s13_m6	0.71	0	0	0	0.41
13	s14_m6	0.69	0	0	0	0.4
14	s15_m6	0.73	-0.01	0	0	0
15	s16_m6	0.53	0.06	0	0	0
16	s17_m6	0.44	0.01	0	0	0
17	s19_m6	0.73	0.04	0	0	0
18	s21_m6	0.67	0	0	0	0.15
19	s23_m6	0.59	0	0	0	0.11
20	s26_m6	0.75	0	0	0	0.21
21	s28_m6	0.55	0	0	0	0.15
22	s29_m6	0.62	0	0.23	0	0
23	s32_m6	0.68	0	0.32	0	0
24	s33_m6	0.64	0	0.31	0	0
25	s34_m6	0.41	0	0.17	0	0
26	s35_m6	0.71	0	0.3	0	0
27	s36_m6	0.53	0	0.26	0	0
28	s37_m6	0.51	0	0.13	0	0
29	s39_m6	0.63	0	0.14	0	0
30	s41_m6	0.76	0	0	0.03	0
31	s42_m6	0.6	0	0	-0.02	0
32	s44_m6	0.68	0	0	-0.05	0
33	s45_m6	0.57	0	0	-0.08	0
34	s46_m6	0.6	0	0	-0.03	0
35	s48_m6	0.52	0	0	0.08	0
36	s50_m6	0.51	0	0	-0.03	0
37	s53_m6	0.49	0	0	0.58	0
38	s54_m6	0.54	0	0	0.69	0

Table C.1 (*continued*)

39	s55_m6	0.55	0	0	0.7	0
40	s56_m6	0.6	0	0	0.65	0
41	s57_m6	0.53	0	0	0.56	0
42	s58_m6	0.41	0	0	0.34	0
43	s59_m6	0.56	0.13	0	0	0
44	s62_m6	0.6	0.15	0	0	0
45	s63_m6	0.27	-0.05	0	0	0
46	s65_m6	0.62	0.14	0	0	0
47	s66_m6	0.54	-0.1	0	0	0
48	s67_m6	0.55	0.07	0	0	0
49	s71_m6	0.59	0	0	0.09	0
50	s72_m6	0.6	0	0	0.02	0
51	s73_m6	0.4	0	0	-0.04	0
52	s74_m6	0.61	0	0	-0.01	0
53	s75_m6	0.74	0.01	0	0	0
54	s76_m6	0.66	0.14	0	0	0
55	s79_m6	0.63	0.01	0	0	0
56	s80_m6	0.61	0.4	0	0	0
57	s81_m6	0.5	0.33	0	0	0
58	s83_m6	0.57	0.15	0	0	0
59	s85_m6	0.48	0.35	0	0	0
60	s86_m6	0.54	0.4	0	0	0

Table C.2 *7th Grade: Factor Loadings of Bifactor Model for Four Standards*

	Item	g	F1	F2	F3	F4
1	s1_m7	0.84	0	0.01	0	0
2	s2_m7	0.61	0	-0.04	0	0
3	s4_m7	0.71	0	0.03	0	0
4	s5_m7	0.64	0	-0.02	0	0
5	s6_m7	0.52	0	0.06	0	0
6	s7_m7	0.75	0	-0.15	0	0
7	s8_m7	0.37	0	0.02	0	0
8	s9_m7	0.72	0	-0.13	0	0
9	s13_m7	0.45	0	0	-0.09	0
10	s16_m7	0.55	0	0	-0.04	0
11	s19_m7	0.67	0	0	0.28	0
12	s20_m7	0.7	0	0	0.61	0
13	s21_m7	0.71	0	0	0.26	0
14	s22_m7	0.68	0	0	0.58	0
15	s24_m7	0.67	0.36	0	0	0
16	s25_m7	0.52	0.37	0	0	0
17	s26_m7	0.66	0.48	0	0	0
18	s27_m7	0.51	-0.03	0	0	0
19	s29_m7	0.72	0	0.02	0	0
20	s31_m7	0.57	0	-0.01	0	0
21	s32_m7	0.57	0	0.12	0	0
22	s33_m7	0.66	0	-0.07	0	0
23	s34_m7	0.39	0	-0.04	0	0
24	s35_m7	0.66	0	0.08	0	0
25	s37_m7	0.48	-0.09	0	0	0
26	s39_m7	0.55	-0.06	0	0	0
27	s42_m7	0.61	0	0.01	0	0
28	s44_m7	0.56	0	-0.1	0	0
29	s46_m7	0.65	0	-0.06	0	0
30	s47_m7	0.52	0	-0.04	0	0
31	s48_m7	0.32	0	0	-0.02	0
32	s49_m7	0.41	0	0	0.12	0
33	s50_m7	0.57	0	0	-0.05	0
34	s52_m7	0.55	0	0	0.25	0
35	s53_m7	0.47	0	0.54	0	0
36	s54_m7	0.57	0	0.68	0	0
37	s55_m7	0.5	0	0.67	0	0
38	s57_m7	0.58	0	0.63	0	0

Table C.2 (*continued*)

39	s58_m7	0.53	-0.01	0	0	0
40	s59_m7	0.5	-0.09	0	0	0
41	s60_m7	0.53	-0.04	0	0	0
42	s62_m7	0.66	-0.13	0	0	0
43	s65_m7	0.52	-0.03	0	0	0
44	s66_m7	0.52	0	0	0.01	0
45	s67_m7	0.69	0	0	0.01	0
46	s68_m7	0.08	0	0	0.08	0
47	s69_m7	0.46	0	0	0.03	0
48	s71_m7	0.55	0	0	0	0
49	s72_m7	0.43	0	0	0.09	0
50	s73_m7	0.72	0	0	0	-0.01
51	s74_m7	0.47	0	0	0	0.01
52	s75_m7	0.46	0	0	0	0.17
53	s76_m7	0.34	0	0	0	0.06
54	s77_m7	0.49	0	0	0	0.07
55	s78_m7	0.5	0	0	0	0.13
56	s79_m7	0.54	0	0	0	0.07
57	s80_m7	0.65	0	0	0	0.57
58	s81_m7	0.62	0	0	0	0.59
59	s83_m7	0.35	0	0	0	0.12
60	s84_m7	0.43	0	0	0	0.09

Table C.3 *8th Grade: Factor Loadings of Bifactor Model for Four Standards*

	Item	g	F1	F2	F3	F4
1	s1_m8	0.64	-0.03	0	0	0
2	s2_m8	0.66	0.54	0	0	0
3	s3_m8	0.61	0.54	0	0	0
4	s4_m8	0.76	0.27	0	0	0
5	s5_m8	0.67	0.03	0	0	0
6	s6_m8	0.41	0.09	0	0	0
7	s7_m8	0.64	0.12	0	0	0
8	s8_m8	0.52	0.07	0	0	0
9	s10_m8	0.54	-0.03	0	0	0
10	s11_m8	0.34	-0.01	0	0	0
11	s12_m8	0.42	0.01	0	0	0
12	s13_m8	0.62	0	0	0	0.62
13	s14_m8	0.52	0	0	0	0.36
14	s15_m8	0.61	0	0	0	0.44
15	s16_m8	0.51	0	0	0	0.42
16	s17_m8	0.54	0	0	0	0.23
17	s18_m8	0.54	0	0	0	0.33
18	s20_m8	0.71	0	0	0.17	0
19	s22_m8	0.72	0	0	0.14	0
20	s23_m8	0.57	0	0	0.15	0
21	s24_m8	0.48	0	0	0	0
22	s26_m8	0.64	0	-0.19	0	0
23	s29_m8	0.69	0	-0.14	0	0
24	s39_m8	0.69	0	0.04	0	0
25	s42_m8	0.67	0	0	0	-0.03
26	s43_m8	0.67	0	0	0	-0.04
27	s44_m8	0.31	0	0	0	-0.09
28	s45_m8	0.65	0	0	0	-0.08
29	s46_m8	0.8	0	0	0	-0.11
30	s47_m8	0.75	0	0	0	-0.13
31	s48_m8	0.49	0	0	0	-0.06
32	s49_m8	0.71	0	0	0	0.02
33	s51_m8	0.65	0	0	-0.08	0
34	s54_m8	0.66	0.01	0	0	0
35	s55_m8	0.65	-0.12	0	0	0
36	s58_m8	0.58	-0.05	0	0	0
37	s60_m8	0.69	0	0	0	-0.15
38	s63_m8	0.62	0	0	0	-0.08

Table C.3 (*continued*)

39	s73_m8	0.75	0	0.16	0	0
40	s74_m8	0.66	0	0.16	0	0
41	s75_m8	0.42	0	-0.03	0	0
42	s76_m8	0.58	0	0.27	0	0
43	s80_m8	0.59	0.09	0	0	0
44	s81_m8	0.6	0.08	0	0	0
45	s82_m8	0.44	0.07	0	0	0
46	s84_m8	0.4	-0.02	0	0	0
47	s86_m8	0.38	0.03	0	0	0
48	s88_m8	0.26	-0.07	0	0	0
49	s89_m8	0.4	0	0	0	0
50	s90_m8	0.7	0	0	0.13	0
51	s91_m8	0.57	0	0	0.54	0
52	s93_m8	0.61	0	0	0.36	0
53	s94_m8	0.49	0	0	0.53	0
54	s95_m8	0.47	0	0	0.09	0
55	s96_m8	0.62	0	0	0.08	0
56	s97_m8	0.48	0	0	0.39	0
57	s98_m8	0.63	0	0.22	0	0
58	s99_m8	0.31	0	0.01	0	0
59	s100_m8	0.44	0	0.3	0	0
60	s101_m8	0.56	0	0.21	0	0

Table C.4 *6th Grade: Factor Loadings of Bifactor Model for Six Benchmarks*

	Item	g	B 1.1	B 1.4	B 2.2	B 3.1	B 3.4	B 4.1
1	s9_m6	0.52	0	0	0	0	0	0.32
2	s10_m6	0.64	0	0	0	0	0	0.43
3	s11_m6	0.46	0	0	0	0	0	0.3
4	s12_m6	0.51	0	0	0	0	0	0.28
5	s13_m6	0.71	0	0	0	0	0	0.42
6	s14_m6	0.69	0	0	0	0	0	0.41
7	s15_m6	0.73	0	-0.09	0	0	0	0
8	s16_m6	0.54	0	-0.05	0	0	0	0
9	s17_m6	0.46	0	-0.02	0	0	0	0
10	s19_m6	0.73	0	-0.08	0	0	0	0
11	s21_m6	0.68	0	0	0	0	0	0.11
12	s23_m6	0.6	0	0	0	0	0	0.07
13	s26_m6	0.77	0	0	0	0	0	0.14
14	s28_m6	0.56	0	0	0	0	0	0.12
15	s33_m6	0.65	0	0	0.21	0	0	0
16	s34_m6	0.4	0	0	0.22	0	0	0
17	s35_m6	0.72	0	0	0.25	0	0	0
18	s36_m6	0.52	0	0	0.36	0	0	0
19	s37_m6	0.51	0	0	0.18	0	0	0
20	s39_m6	0.63	0	0	0.12	0	0	0
21	s41_m6	0.73	0	0	0	0.29	0	0
22	s42_m6	0.55	0	0	0	0.37	0	0
23	s44_m6	0.64	0	0	0	0.22	0	0
24	s45_m6	0.52	0	0	0	0.48	0	0
25	s46_m6	0.51	0	0	0	0.74	0	0
26	s53_m6	0.49	0	0	0	0	0.59	0
27	s54_m6	0.53	0	0	0	0	0.7	0
28	s55_m6	0.53	0	0	0	0	0.71	0
29	s56_m6	0.6	0	0	0	0	0.65	0
30	s57_m6	0.52	0	0	0	0	0.57	0
31	s58_m6	0.41	0	0	0	0	0.34	0
32	s65_m6	0.63	0.13	0	0	0	0	0
33	s66_m6	0.5	0.19	0	0	0	0	0
34	s67_m6	0.54	0	0	0	0	0	0
35	s75_m6	0.72	0.25	0	0	0	0	0
36	s76_m6	0.67	0.38	0	0	0	0	0
37	s79_m6	0.63	0	-0.05	0	0	0	0
38	s80_m6	0.62	0	0.37	0	0	0	0

Table C.4 (*continued*)

39	s81_m6	0.51	0	0.27	0	0	0	0
40	s83_m6	0.58	0	0.13	0	0	0	0
41	s85_m6	0.5	0	0.31	0	0	0	0
42	s86_m6	0.55	0	0.44	0	0	0	0

Table C.5 *7th Grade: Factor Loadings of Bifactor Model for Four Benchmarks*

	Item	g	B 1.1	B 1.4	B 2.2	B 3.1
1	s1_m7	0.78	0	0	0.41	0
2	s2_m7	0.57	0	0	0.41	0
3	s4_m7	0.64	0	0	0.54	0
4	s5_m7	0.6	0	0	0.54	0
5	s24_m7	0.66	0.36	0	0	0
6	s25_m7	0.5	0.38	0	0	0
7	s26_m7	0.62	0.56	0	0	0
8	s27_m7	0.51	-0.02	0	0	0
9	s29_m7	0.69	0	0	0.09	0
10	s31_m7	0.56	0	0	-0.05	0
11	s32_m7	0.56	0	0	0.25	0
12	s33_m7	0.66	0	0	-0.05	0
13	s34_m7	0.4	0	0	-0.02	0
14	s35_m7	0.64	0	0	0.2	0
15	s37_m7	0.48	0	0.1	0	0
16	s39_m7	0.55	0	0.1	0	0
17	s42_m7	0.61	0	0	-0.12	0
18	s44_m7	0.56	0	0	0.02	0
19	s46_m7	0.67	0	0	-0.14	0
20	s47_m7	0.52	0	0	0.02	0
21	s58_m7	0.53	0	0.05	0	0
22	s59_m7	0.49	0	0.64	0	0
23	s60_m7	0.55	0	0.03	0	0
24	s62_m7	0.65	0	0.28	0	0
25	s65_m7	0.53	0	0	0	0
26	s66_m7	0.48	0	0	0	0.21
27	s67_m7	0.68	0	0	0	-0.03
28	s68_m7	0.05	0	0	0	0.12
29	s69_m7	0.45	0	0	0	0.2
30	s71_m7	0.53	0	0	0	0.1
31	s72_m7	0.42	0	0	0	0.58

Table C.6 *8th Grade: Factor Loadings of Bifactor Model for Six Benchmarks*

	Item	g	B 1.1	B 1.4	B 2.2	B 3.1	B 3.4	B 4.1
1	s7_m8	0.64	0.05	0	0	0	**0**	0
2	s8_m8	0.52	0.23	0	0	0	**0**	0
3	s10_m8	0.54	0.11	0	0	0	**0**	0
4	s11_m8	0.34	0.04	0	0	0	**0**	0
5	s12_m8	0.41	0.61	0	0	0	**0**	0
6	s20_m8	0.66	0	0	0	0.61	**0**	0
7	s22_m8	0.68	0	0	0	0.62	**0**	0
8	s23_m8	0.54	0	0	0	0.37	**0**	0
9	s39_m8	0.67	0	0	0.07	0	**0**	0
10	s42_m8	0.67	0	0	0	0	**0**	-0.08
11	s43_m8	0.66	0	0	0	0	**0**	0.16
12	s44_m8	0.31	0	0	0	0	**0**	0.1
13	s45_m8	0.65	0	0	0	0	**0**	0.05
14	s46_m8	0.81	0	0	0	0	**0**	0.14
15	s47_m8	0.76	0	0	0	0	**0**	0.01
16	s48_m8	0.5	0	0	0	0	**0**	0.02
17	s49_m8	0.72	0	0	0	0	**0**	-0.09
18	s51_m8	0.67	0	0	0	-0.08	**0**	0
19	s54_m8	0.66	0	0.01	0	0	**0**	0
20	s55_m8	0.66	0	0	0	0	**0**	0
21	s58_m8	0.59	0	0.06	0	0	**0**	0
22	s60_m8	0.66	0	0	0	0	**0**	0.5
23	s63_m8	0.58	0	0	0	0	**0**	0.73
24	s73_m8	0.75	0	0	0.23	0	**0**	0
25	s74_m8	0.65	0	0	0.48	0	**0**	0
26	s75_m8	0.39	0	0	0.12	0	**0**	0
27	s76_m8	0.59	0	0	0.12	0	**0**	0
28	s80_m8	0.58	0	0.2	0	0	**0**	0
29	s81_m8	0.6	0	0.18	0	0	**0**	0
30	s82_m8	0.45	0	0.7	0	0	**0**	0
31	s84_m8	0.41	0	0.09	0	0	**0**	0
32	s90_m8	0.71	0	0	0	0	**0.12**	0
33	s91_m8	0.55	0	0	0	0	**0.59**	0
34	s93_m8	0.6	0	0	0	0	**0.38**	0
35	s94_m8	0.48	0	0	0	0	**0.55**	0
36	s95_m8	0.46	0	0	0	0	**0.09**	0
37	s96_m8	0.63	0	0	0	0	**0.07**	0
38	s97_m8	0.48	0	0	0	0	**0.39**	0

APPENDIX D: NUMBER OF EXAMINEES

FOR GENDER X SCHOOL LUNCH PROGRAM AND

RACE X SCHOOL LUNCH PROGRAM

Table D.1 *Frequencies of Gender X School Lunch Program*

		School lunch program			Total
		Regular price	Reduced price	Free	
Female	Count	811	140	409	1360
	% within gender	59.6%	10.3%	30.1%	100%
	% within lunch	51.3%	46.7%	52.0%	51.0%
Male	Count	769	160	378	1307
	% within gender	58.8%	12.2%	28.9%	100%
	% within lunch	48.7%	53.3%	48.0%	49.0%
Total	Count	1580	300	787	2667
	% within gender	59.2%	11.2%	29.5%	100%
	% within lunch	100%	100%	100%	100%

Table D.2 *Frequencies of Race X School Lunch Program*

		lunch_m6			Total
		Regular price	Reduced price	Free	
American Indian	Count	17	6	21	44
	% within race	38.6%	13.6%	47.7%	100%
	% within lunch	1.1%	2.0%	2.7%	1.6%
Asian	Count	54	9	20	83
	% within race	65.1%	10.8%	24.1%	100%
	% within lunch	3.4%	3.0%	2.5%	3.1%
Black	Count	66	39	166	271
	% within race	24.4%	14.4%	61.3%	100%
	% within lunch	4.2%	13.0%	21.1%	10.2%
Hispanic	Count	79	55	235	369
	% within race	21.4%	14.9%	63.7%	100%
	% within lunch	5.0%	18.3%	29.9%	13.8%
White	Count	1325	177	314	1816
	% within race	73.0%	9.7%	17.3%	100%
	% within lunch	83.9%	59.0%	39.9%	68.1%
Multi or Missing	Count	39	14	31	84
	% within race	46.4%	16.7%	36.9%	100%
	% within lunch	2.5%	4.7%	3.9%	3.1%
Total	Count	1580	300	787	2667
	% within race	59.2%	11.2%	29.5%	100%
	% within lunch	100%	100%	100%	100%

APPENDIX E

MATHEMATICAL ACHIEVEMENT CHANGE FROM 6TH TO 8TH GRADE FOR EACH OF THE FOUR GROUPS (FEMALE, MALE, REGULAR PRICE LUNCH, AND FREE/REDUCED PRICE LUNCH)

Table E. 1 *Female: Average Mathematical Achievement Change from 6th to 8th grade*

Content	Trait level at 6th grade (θ_1)	θ_2	θ_3	Trait level at 7th grade ($=\theta_1+\theta_2$)	Trait level at 8th grade ($=\theta_1+\theta_2+\theta_3$)
Standard 1	0.198	0.940	0.064	1.137	1.201
Standard 2	0.267	0.930	0.126	1.197	1.323
Standard 3	0.170	0.684	0.506	0.853	1.359
Standard 4	0.585	0.223	0.225	0.808	1.033
Benchmark 1.1	0.160	0.637	0.241	0.797	0.907
Benchmark 1.4	0.140	0.205	0.227	0.345	0.572
Benchmark 2.2	0.140	0.205	0.227	0.345	0.572
Benchmark 3.1	0.159	0.646	0.187	0.805	0.992
Benchmark 3.4	0.320		0.673		0.993
Benchmark 4.1	0.229		0.810		1.039

Table E.2 *Male: Average Mathematical Achievement Change from 6th to 8th grade*

Content	Trait level at 6th grade (θ_1)	θ_2	θ_3	Trait level at 7th grade ($=\theta_1+\theta_2$)	Trait level at 8th grade ($=\theta_1+\theta_2+\theta_3$)
Standard 1	0.222	0.832	0.089	1.054	1.143
Standard 2	0.237	0.813	0.130	1.050	1.180
Standard 3	0.115	0.601	0.534	0.716	1.250
Standard 4	0.486	0.310	0.233	0.796	1.029
Benchmark 1.1	0.210	0.580	0.315	0.790	0.930
Benchmark 1.4	0.102	0.213	0.220	0.315	0.535
Benchmark 2.2	0.102	0.213	0.220	0.315	0.535
Benchmark 3.1	0.176	0.599	0.106	0.775	0.882
Benchmark 3.4	0.251		0.621		0.872
Benchmark 4.1	0.155		0.747		0.902

Table E.3 *Regular Price Lunch: Average Mathematical Achievement Change from 6th to 8th grade*

Content	Trait level at 6th grade (θ_1)	θ_2	θ_3	Trait level at 7th grade ($=\theta_1+\theta_2$)	Trait level at 8th grade ($=\theta_1+\theta_2+\theta_3$)
Standard 1	0.447	0.921	0.140	1.368	1.508
Standard 2	0.461	0.941	0.176	1.402	1.578
Standard 3	0.326	0.726	0.572	1.051	1.623
Standard 4	0.712	0.254	0.326	0.966	1.292
Benchmark 1.1	0.328	0.684	0.377	1.012	1.132
Benchmark 1.4	0.315	0.264	0.284	0.580	0.864
Benchmark 2.2	0.367	1.052	0.243	1.419	1.662
Benchmark 3.1	0.272	0.699	0.234	0.970	1.205
Benchmark 3.4	0.382		0.771		1.153
Benchmark 4.1	0.351		0.902		1.254

Table E.4 *Free/Reduced Price Lunch: Average Mathematical Achievement Change from 6th to 8th grade*

Content	Trait level at 6th grade (θ_1)	θ_2	θ_3	Trait level at 7th grade ($=\theta_1+\theta_2$)	Trait level at 8th grade ($=\theta_1+\theta_2+\theta_3$)
Standard 1	-0.135	0.837	-0.016	0.702	0.686
Standard 2	-0.050	0.772	0.058	0.722	0.780
Standard 3	-0.122	0.523	0.444	0.400	0.844
Standard 4	0.280	0.284	0.089	0.564	0.653
Benchmark 1.1	-0.024	0.500	0.132	0.476	0.608
Benchmark 1.4	-0.161	0.128	0.136	-0.033	0.104
Benchmark 2.2	-0.094	0.913	0.054	0.819	0.873
Benchmark 3.1	0.016	0.513	0.021	0.529	0.550
Benchmark 3.4	0.147		0.469		0.616
Benchmark 4.1	-0.037		0.600		0.563

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Ansell, E. & Doerr, H. M. (2000). NAEP findings regarding gender:Achievement, affect, and instructional experiences. In E. A. Silver & P. A. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 73-106). Reston, VA: National Council of Teachers of Mathematics, Incorporated.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D. (1997). The nominal categories model. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-50). N.Y.: Springer.

Case, R. (1985). *Intellectual development: Birth to adulthood*. Orlando, FL: Academic Press.

Case, R., & Okamoto, Y., in collaboration with Griffin, S., McKeough, A., Bleiker, C., Henderson, B., & Stephenson, K. M. (1996). The role of central conceptual structures in the development of children's thought. Monographs of the society for research in child Development, 61 (1-2, serial No. 246).

Crane, J. (1996). Effects of home environment, SES, and maternal test scores on mathematics achievement. *Journal of Educational Research, 89*, 305-315.

Du Toit, M. (2003). *IRT from SSI* . Lincolnwood, IL : SSI Scientific Software international. Inc.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49,* 175-186.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56(3),* 495-516.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests.* (pp. 125-150). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Embretson, S. E. (1995). A measurement model for linking individual change to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement, 32,* 277-294.

Embretson, S. E. (1997a). Multicomponent latent trait models. In W. van der Linden & R. Hambleton, *Handbook of modern item response theory*. New York: Springer-Verlag, pp. 305-322.

Embretson, S. E. (1997b). Structured ability models in tests designed from cognitive theory. In M. Wilson, G. Engelhard & K. Draney (Eds.) *Objective Measurement III* (pp. 223-236). Norwood, NJ: Ablex.

Embretson, S. E. (2000). Multidimensional measurement from dynamic tests: Abstract reasoning under stress. *Multivariate Behavioral research, 35 (4),* 505-543.

Embretson, S. E. (2003). Cognitive models for psychometric properties of GRE quantitative items. *ETS Progress Report for Research Project.*

Embretson, S. E. (2006). *Cognitive models for the psychometric properties of GRE quantitative items.* Final Report. Educational Testing Service.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*.

Embretson, S. E., & McCollam, K. M. (2000). A multicomponent rasch model for measuring covert processes. In M. Wilson & G. Engelhard (Eds.), *Objective Measurement V5* (pp. 203-218). Norwood, NJ: Ablex.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. E., & Yang, X. (2006). Multicomponent latent trait models for complex tasks. *Journal of Applied Measurement, 7 (3),* 540-557.

Embretson, S. E., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26, Psychometrics* (pp.747-767). American: Elsevier.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fisher, G. H. (1995). Some neglected problems in IRT. *Psychometrika,* 60, 459-487.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 221–224). New York: Springer.

Fischer, G. H., & Formann, A. K. (1982). Same applications of logistic latent trait models with linear constraints on the parameters. Applied Psychological Measurement, *6,* 397-416.

Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrica, 59,* 177-192.

Fischer, G. H., & Ponocny-Seliger, E. (1998). Structural Rasch modeling. Handbook of the usage of LPCM-Win 1.0. Groningen: PROGAMMA.

Fennema, E. (1996). Mathematics, gender and research. In G. Hanna (Ed), *Towards gender equity in mathematics education* (pp. 9-26). Norwell, MA: Kluwer Academic Publishers.

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research, 59,*185-213.

Glas, C. A.W., & Van der Linden,W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247-261.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: application and limitations of four different approaches. Method of Psychological Research Online, 2(1), 1-18.

Gray, M. (1996). Gender and mathematics: Mythology and misogyny. In G. Hanna (Ed), *Towards gender equity in mathematics education* (pp. 27-38). Norwell, MA: Kluwer Academic Publishers.

Guo, G. (1998). The timing of the influences of cumulative poverty on children's cognitive ability and achievement. *Social Forces, 777,* 257-288.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury park, California: Sage Publications, Inc.

Hanna, G. (2003). Reaching gender equity in mathematics education. *The Education Forum, 67, 204-214.*

Hopkins, T. M. (2004). *Gender Issues in Mathematics Achievement in Tennessee:Does Rural School Locale Matter?* Unpublished doctoral dissertation, University of

Tennessee at Knoxville.Katz, I.R., Bennett, R.E., & Berger, A.E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*(1), 39-57.

Kansas State Department of Education (2003). *Kansas Mathematical Standards.* Available at www.ksde.org/Default.aspx?tabid=156

Kelderman, H., & Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Kuenzi, J. J. (2008). *CRS Report for Congress: Science, Technology, Engineering, and Mathematics (STEM) Education: Background, Federal Policy, and Legislative Action*. Domestic Social Policy Division

Leahy, E. & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces 80*, 713-733.

Mckinley, R. L., & Reckase, M.D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Rep. ONR 82-1). Iowa City IA: American College Testing.

Mayer, R., Lakin, J., & Kadane, P. (1984). A cognitive analysis of mathematical problem solving. In R. Sternberg (Ed), *Advances in the psychology of human intelligence (Vol.2).* Hillsdale, NJ: Erlbaum.

Muraki, E., & Engelhard, G. (1985).Full information item factor analysis: Application of EAP score. *Applied Psychological Measurement, 9*, 417-430.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithem. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). N.Y.: Springer.

Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: Attitude and stereotype threat susceptibility in Italian children. *Developmental Psychology, 43(3),* 747-759.

National Mathematics Advisory Panel. (2008). *Reports of the Task Groups and Subcommittees.* Washington, D.C.: U.S. Department of Education.

Newstead, S. E., Bradon, P., Handley, S. J., Dennis, I., & Evans, J. S. B. T. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning, 12(1),* 62-90.

Ormrod, J. E. (2008). *Human Learning* (5th Edition ed.). New Jersey: Pearson Prentice Hall.

Piaget, J., & Garcia, R. (1989). *Psychogenesis and the history of science*. New York: Columbia University Press.

Piaget, J., & Inhelder, B. (1967).*The child's conception of space* (F. J. Langdon & J.L. Lunzer, Trans). New York: Norton. (Original work published 1948)

Piaget, J., & Inhelder, B. (1969).*The Psychology of the Child* (H. Weaver, Trans.). New York: Basic Books.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Rost, J. (1990). Rash models in latent classes: An integration of two approaches to item a nalysis. *Applied Psychological Measurement, 14*, 271-282.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Iowa City, IA: Psychometric Society.

Samejima, F. (1997). Graded response model. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). N.Y.: Springer.

Shepard, L. A. (2008) Commentary on the National mathematics advisory panel recommendations on assessment. *Educational Researcher, 37(9),* 602-609.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75*, 428-444.

Siegler, R. S., & Ramani, G. B. (2008). Playing board games promotes low-income children's numerical development. *Developmental Science, 11:5*,655-661.

Spada, H., & McGaw, B. (1985) The assessment of learning effects with linear logistic test models. In Embretson, S. E. (Ed.). *Test design: Developments in psychology and psychometric* (pp. 169-193). New York: Academic Press.

Sprigler, D. M, & Alsup, J. K. (2003). An analysis of gender and mathematical reasoning ability sub-skill of analysis-sythesis. *Education 123*, 763-769.

Tatsuoka, K.K., Corter, J.E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*, 901-926.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one parameter logistic model. *Psychometrika, 47*, 201-214.

Thissen, D. (2009). The MEDPRO project: An SBIR project for a comprehensive IRT and CAT software system-IRT software. In D.J. Wesis (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Thissen, D. (2010). IRTPRO Beta Features and Operation [Computer software manual], IL: Scientific Software International.

Thissen, D., Cai, L., & Bock, R.D. (in press). The nominal item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications*.

Traub, R.E., & Fisher, C.W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*(3), 355-369.

von Davier, M., Xu, X., & Carstensen, C. H. (2009). *Using the general diagnostic model to measure learning and change in a longitudinal large-scale assessment* (ETS Research Report RR- 09-28). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service).

Weiss, D. J. & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D.J. Wesis (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*.

Wellesly College (1992). How schools shortchange girls. A study of major findings on girls and education. The AAUW Report. American Association of University Women Educational Foundation, Washington, D. C. (Eric Document Reproduction Service No. ED 339674)

Wilson, M. (1985). Measuring stages of growth: A psychometric model of hierarchical development. Occasional paper No. 19. Hawthorn, Victoria: Australian Council for Educational Research.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.